Genetic networks of antibacterial responses of eukaryotic cells. Bioinformatics analysis and modeling

Vom Fachbereich für Biowissenschaften und Psychologie
der Technischen Universität Carolo-Wilhelmina
zu Braunschweig

zur Erlangung des Grades einer
Doktorin der Naturwissenschaften
(Dr.rer.nat.)

genehmigte

D i s s e r t a t i o n

von **Ekaterina Shelest**
aus Novosibirsk

1. Referent: Prof. Dr. D. Jahn

2. Referent: Prof. Dr. E. Wingender.

eingereicht am: 23.11.2005

mündliche Prüfung (Disputation) am: 7.02.2006

2006

(Druckjahr)

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Gemeinsamen Naturwissenschaflichen Fakultät, vertreten durch den Mentor die Betreuerin der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Shelest, E., Kel, A.E., Goessling, E. & Wingender, E. Prediction of potential C/EBP/NF-kappaB composite elements using matrix-based search methods. In Silico Biol. 3: 71-79. (2003).

Shelest, E. & Wingender, E. Construction of predictive promoter models on the example of antibacterial response of human epithelial cells. Theor Biol Med Model. 2(1):2. (2005).

Tagungsbeiträge

Shelest, E., Kel-Margoulis, O., Kel, A. & Wingender, E.: Bioinformatics representation of cellular responses to bacterial infection. (Vortrag). Cell signaling, transcription and translation as therapeutic targets. Luxembourg (2002).

Shelest, E., Kel, A.E., Gößling, E. & Wingender E.: Prediction of potential C/EBP/NF-kappaB composite elements using matrix-based search methods. (Vortrag). The 3rd International Conference on Bioinformatics of Genome Regulation and Structure, BGRS 2002, Novosibirsk, Russia (2002).

Shelest, E., Kel, A., Gößling, E. & Wingender, E.: Composing a promoter model for antibacterial response of epithelial cells. (Poster). European Conference on Computational Biology, ECCB 2002, Saarbruecken, Germany (2002).

Shelest E., Kel A.E. & Wingender E.: Constructing a promoter model for antibacterial response of lung epithelial cells. (Poster). Gordon Research Conference "Bioinformatics: from predictive models to inference". Oxford UK (2003).

Shelest, E., Kel, A.E. & Wingender, E.: Constructing a promoter model for antibacterial response of lung epithelial cells. (Poster). European Conference on Computational Biology, ECCB 2003, Paris, France (2003).

Shelest E., Sauer T. & Wingender E.: Regulatory networks of antibacterial response. (Poster). Symposium NGFN, Tübingen, Germany (2003).

Shelest E. & Wingender E.: Identification of immune-related target genes by application of a

promoter model. (Vortrag). 6th EMBL Transcription Meeting, Heidelberg, Germany (2004).

Shelest, E. & Wingender, E.: Investigation of distances in transcription factor binding site pairs. (Poster). European Conference on Computational Biology, ECCB 2005, Madrid, Spain. (2005)

All models are wrong—but some are useful.

George E.P. Box, 1979

Knowledge of some principles easily compensates for the

ignorance of some facts.

Claude Adrien Helvetius

(1715-1771)

CONTENTS

List of Abbreviations

asialoGM1 - asialoganglioside 1

CE - composite element

GRN – gene regulatory network

IFN - interferon

IRF-3 - interferon regulatory factor-3

LPS - lipopolysaccharide

MALP-2 - macrophage activating lipopeptide of 2 kDa

MyD88 - myeloid differentiation factor 88

PAI-1 - *Pseudomonas* autoinducer 1

PAMP - pathogen-associated molecular pattern

PCN - pyocyanin

PMN - polymorphonuclear leukocytes

PRR - pattern recognition receptor

PWM – positional weight matrix

SP – signaling pathways

STP - signal transduction pathways

TIR – Toll/IL-1 receptor domain

TF – transcription factor

TFBS - transcription factor binding site

TLR - Toll-like receptor

TN – transcription network

1. INTRODUCTION

Over the past several years, rapid development of new powerful methods, like expression profiles using cDNA arrays, supplied researchers with a vast quantity of gene expression data. In parallel, functional knowledge of individual genes encoding components of the cell signaling, metabolic and regulatory pathways has been accumulated through the years. These data not only contain information about how, why and what is expressed in cells under different conditions, but demand for appropriate methods to extract this information. Now the time is coming, when quantity (of knowledge) can be transformed into quality (of knowledge). The amount of information, on the one hand, and development of computational approaches on the other enable us to analyze this information. This analysis can be conducted in two directions: in the direction of a general overview of the processes happening in and between the cells, and in the direction of understanding the characteristic details of these processes.

This work is devoted to the construction of promoter models as one of necessary steps required for the construction of regulatory networks. It includes the development of methods of promoter model construction and the application of these methods to the description in terms of transcription regulation of several defensive eukaryotic systems. In this introduction I would like to give a short overview of the investigated systems, as well as of the methods and problems of promoter model construction.

1.1. Gene regulatory networks, transcription networks and promoter models

1.1.1. Biological networks

In molecular biology, we may distinguish between different kinds of networks and pathways: metabolic, protein-protein interaction and regulatory.

Metabolic networks represent the conversion of metabolites among each other, by complex networks of enzymatically catalyzed reactions. These networks can be viewed in a metabolite- or an enzyme-centric way.

Protein-protein-interaction networks unspecifically represent the network of physical, sometimes just functional interactions of the components of a cell's or system's proteome.

Regulatory networks comprise different kinds of the following networks: gene regulatory; transcription (which are subnetworks of gene regulatory networks); signaling and signal

transduction, the latter may be considered as specification of signaling networks.

Gene regulatory networks (GRN) describe the functional interaction of genes through their products, which may influence the activity of other genes and/or their products. As can be assumed from this vague description, GRNs in this broad sense represent rather unspecific, usually phenomenologically characterized interactions without specific statement of the underlying mechanisms.

The process of gene regulation depends largely on a special group of proteins known as transcription factors. Being proteins themselves, they can be also the subjects of gene interaction, their expression being dependent on other genes or/and signaling pathways of the cells triggered under some circumstances. The supression or activation of mRNA synthesis of a gene depends on the association of active transcription factors with regulatory sequence(s) of the corresponding gene. One transcription factor may be responsible for the activation of a large number of genes; some of them may be other transcription factors. Moreover, the transcription factors tend to cooperate in activating gene expression, so the interaction network expands. Finally, the gene regulatory network determines which genes are expressed under the given conditions and to which extent, how the cell will response to the diverse environmental irritants and changes, how will it develop and change following the intracellular signals. Transcription factors together with the regulated genes constitute a subnetwork of the gene regulatory network, which is called transcription network (TN).

Signaling networks or pathways (SP) represent the transfer of environmental signals to defined target compartments or molecules inside the cell. The term is mostly used synonymously with signal transduction pathways (STP), though the latter is frequently used more specifically to describe the transduction of extracellular signals towards the nucleus, to alter the genetic program of the target cell. This would exclude the regulation of metabolic enzymes or structural components of a cell in response to such signals.

In comparison with transcription networks, SP describe the regulatory processes in a broader sense. The TNs can be in a large part overlapping with SPs with the exception of those transcription factors, which are not a subject of signal transduction (such as Sp1 transcription factor which is transcribed at a steady level and practically does not depend on the signals triggering the other transcription factors which participate in the same transcription network). The overlap of TNs with STPs can occur in much smaller part (only in the description of transcription factors and their transmission to nucleus).

1.1.2. Definition of a promoter model

A promoter is a region of DNA, which directs RNA polymerase binding before initiating the transcription of DNA into RNA. Among the three eukaryotic RNA polymerases, RNA polymerase II recognizes many thousands of promoters of protein-coding genes. Most of these promoters have the Goldberg-Hogness or TATA box that is centered around position -30 and has the consensus sequence 5'-TATAAAA-3'. Some eukaryotic genes also contain an initiator element (Inr). Most naturally occurring initiator elements have a cytosine (C) at the −1 position and an adenine (A) residue at the transcription-start site (+1). Several promoters have a CAAT box around -90 with the consensus sequence 5'-GGCCAATCT-3'. Transcription of genes with promoters containing a TATA box or initiator element begins at a well-defined initiation site. However, transcription of many protein-coding genes has been shown to begin at any one of multiple possible sites over an extended region, often 20 – 200 base pairs in length. As a result, such genes give rise to mRNAs with multiple alternative 5′ ends. These genes, which generally are transcribed at low rates (e.g., genes encoding the enzymes of intermediary metabolism, often called "housekeeping genes"), do not contain a TATA box or an initiator. Most genes of this type contain a CG-rich stretch of 20 – 50 nucleotides within ≈100 base pairs upstream of the start-site region. Transcription by polymerase II is also affected by more distant elements, known as enhancers.

Eukaryotic promoters can be discussed in two ways:

- Considering only the TATA-box or initiator sequences that determine the initiation site in the template and using the term "promoter-proximal elements" for control regions lying within 100 – 200 base pairs upstream of the start site. This is usually referred to as "core promoter" in the range of approximately –50 to +10, thus basically comprising the transcription initiation site and the TATA box around –30.
- Considering the other, more specific transcription factors and their binding sites, which are responsible for transcription initiation in tissue-, cell type- or developmental stage-specific manner. In such consideration, the length of the promoter region may be much larger, but it is normally considered that it should not exceed several hundreds of base pairs upstream the transcription start site (Davuluri *et al.,* 2001).

Consequently, the term „promoter model" can have two meanings:

- a set of characteristic features of promoters as a class of sequences; we can call it "general promoter models";
- a set of characteristic features of specific promoters, e.g., promoters of co-regulated genes, or genes with a common function, etc.; this we can call "specific promoter

models".

In this work we consider only the promoter models in the second meaning; thus, we will call them further on simply "promoter models".

Definition: A promoter model is a combination of sequence elements modulating transcription and characterizing promoter(s) of a certain gene/group of genes (Bailey and Noble, 2003). By sequence elements we mean in this case transcription factor binding sites, although this term may be understood in a broader sense.

Transcription factors (TFs) form an own functional class of proteins. They can be defined as proteins, which, after nuclear translocation, regulate transcription by stoichiometric interaction with specific DNA sequences or with proteins that are specifically bound to DNA (Wingender, 1997). In this sense, and corresponding to the two different meanings of the term "promoter", we may differentiate between general TFs (such as TFIID, TFIIB, etc.) which necessarily work more or less with all promoters, and specific ones (or "upstream factors"), which act via specific sequence elements (*cis*-acting elements). In this work, we deal exclusively with promoter and TFs in the second sense, i. e. we do not consider modeling of general promoter features with the purpose of general promoter prediction (e.g., Scherf *et al.,* 2000; Davaluri *et al.,* 2001; Bajic *et al.,* 2002). It is a widely accepted opinion that transcription factors, which are involved in a certain cellular response, tend to cooperate and act in most cases in a synergistic manner (Brazma *et al.,* 1998, Fickett and Wasserman, 2000, Wagner, 1999, Werner *et al.,* 2003). Therefore, their binding sites are organized in a non-random manner and co-regulation of genes entails the existence of a characteristic combination of TFBS in their promoters.

Promoter models are constructed based on the investigation of some true positive examples (sets of promoters of co-expressed or co-regulated genes).

Promoter models may be used for at least two purposes: (i) having a promoter model for a certain set of genes, we can search for other genes possessing the same combination, thus being potentially involved in the same cellular response (predictive models); (ii) understanding the structure of promoters, we can derive the information about the ascending pathways which have triggered the investigated set of genes (descriptive models). Having at hand the results of microarray analyses, we can investigate the promoters of co-expressed genes and with the help of promoter models define or refine the regulatory networks of the corresponding processes. The successful development of the methods of promoter model construction is, therefore, an important approach of systems biology.

1.2. Biological systems addressed

The methods of promoter model construction developed in this work were applied to the description of several defensive eukaryotic systems, all of which belong to the innate immune responses. Thus, in the following paragraphs I introduce some principles of innate immunity, and, more specifically, of causes and pathways of antibacterial responses.

1.2.1. Antibacterial response: Innate immunity

The ability of a quick response to infecting pathogens is essential to the survival of any organism. In mammals, the immune system consists of two subsystems: innate and adaptive immunities. The first is an ancient mechanism which is shared among most multicellular organisms, whereas the adaptive immunity is a novelty which evolved on the latest steps of the evolution and is found only in the highest vertebrates (Hoffmann *et al.,* 1999). Adaptive immune responses are required for complete clearance of many pathogens, but are not effective during early stages of the infection, because antigen-specific lymphocytes require several days of clonal expansion to reach sufficient numbers (Whitsett *et al.,* 2004). The innate immune system, on the contrary, responds almost immediately. It limits pathogen spreading until the adaptive immune system becomes effective. Thus, the complex immune response of mammals relies on the communication between both immune systems. The innate immune system is the first line of defense against infectious microorganisms. It provides initial and rapid host defense, mediates inflammatory responses and, thus, has a profound impact on the establishment of adaptive immune responses (Hoffmann *et al.,* 1999; Takeuchi and Akira, 2001).

The basic mechanisms of innate immunity are conserved and found in most animals (beginning with insects) and even in plants (Hoffmann *et al.,* 1999; Silverman and Maniatis, 2001). The main characteristic of the innate immune system is the usage of germline-encoded pattern recognition receptors to recognize the invading pathogens (Medzhitov and Janeway, 1997). The prerequisite for this is the existence of common structural features shared by different microorganisms in spite of a tremendous variety of the latter; the examples of these structures are lipopolysaccharide (LPS) from Gram-negative bacteria, lipoteichoic acid (LTA) from Gram-positive bacteria, lipoarabinomannan (LAM) from mycobacterium, unmethylated DNA and bacterial lipoproteins (Takeuchi and Akira, 2001). These invariant bacterial structures are called pathogen-associated molecular patterns (PAMPs). The important feature of PAMPs is that they are not expressed by hosts (Medzhitov and Janeway, 1997).

The mechanisms of innate immunity are well described and reviewed in many papers

(Medzhitov and Janeway, 1997; Hoffmann *et al.,* 1999; Takeuchi and Akira, 2001; Zhang and Ghosh, 2001, and many others). To give a short overview, recognition of microbial products by receptors on effector cells results in the induction of pro-inflammatory cytokines and activation of the inflammatory response. The effector cells can be neutrophils, monocytes, macrophages and endothelial and mucosal epithelial cells. The PAMPs of invading pathogens are recognized by pattern recognition receptors (PRRs). The examples of PRRs include CD14, β2-integrins (CD11/CD18), C-type lectins, macrophage scavenger receptors, and complement receptors (CR1/CD35, CR2/CD21) (Medzhitov and Janeway, 1997). These PRRs are expressed as either membrane-bound or soluble proteins.

Recognition of PAMPs by PRRs results in the activation of different intracellular signaling cascades that in turn lead to the expression of various effector molecules (Medzhitov and Janeway, 1997). These effector molecules can be divided into three groups:

1. various antimicrobial peptides as well as reactive oxygen and nitrogen intermediates;
2. cytokines, chemokines, adhesion molecules, and acute phase proteins;
3. co-stimulatory molecules B7.1 and B7.2.

The first two groups are responsible for the early host defense, the first providing microbicidal activity and immediate protection for hosts and the second being involved in inflammation and as well as the development of adaptive immune responses; the molecules of the third group bind CD28 on T cells and act as the second signal for T-cell activation. Therefore, signaling by the PRRs helps to bridge innate and adaptive immunity and allows the host to cope more efficiently with microbial infection.

Since the specific focus of this work was initially concentrated on the response of human lung epithelial cells on the binding of *P. aeruginosa*, I will give a short description of this bacterium and its specific characteristics important for triggering the antibacterial response.

1.2.2. *Pseudomonas aeruginosa*

1.2.2.1. General characteristics, virulence and biofilm formation

Pseudomonas aeruginosa is a Gram-negative, free-living aerobic bacterium prospering in most environments, and has been isolated from soil, water, and sewage. Being an opportunistic pathogen, it rarely infects uncompromised tissues and is mostly discovered in patients with impairments in their host defence. The predisposing conditions can be leukemia, lymphoma, cystic fibrosis (CF), diffuse panbronchiolitis (DPB), AIDS, and burns. *P. aeruginosa* is the most common cause of ICU(intensive care unit)-associated pneumonia and

is responsible for approximately 10% of the 2 million nosocomial infections that occur annually (Wozhniak and Keyser, 2004).

The main characteristics of *P. aeruginosa* are multiple virulence factors and natural resistance to treatment with many antibiotics. The latter makes it especially hard to treat the *Pseudomonas*-caused pneumonia. The virulence factors are critical in the establishment of acute as well as chronic infections; they include LPS, flagella, pili, quorum-sensing molecules, proteases, toxins and others. This repertoire of virulence factors promotes adherence to the host cells, damages host tissues, elicits inflammations, and possibly disrupts host defences by altering gene expression in host cells (Cobb *et al.,* 2004).

The antibiotic resistance occurs mainly due to biofilm formation. *P. aeruginosa* can exist in two forms: planktonic and biofilm. In its planktonic form, *P. aeruginosa* is a free-swimming cell that moves by means of a single polar flagellum. In its sessile biofilm form, the bacteria attach to abiotic surfaces or organic substrates. Biofilms protect bacteria from phagocytosis, antibiotics, ciliary action of the respiratory tract, opsonizing antibodies, and complement (Costerton *et al.,* 1999). It is estimated that biofilms account for approximately 60% of microbial infections in the body (Costerton *et al.,* 1999; Mah and O'Toole, 2001). The bacteria in old biofilms can stand 100-1000-fold higher minimal antibiotics concentrations than the planctonic forms, whereas the young biofilms are less resistant (Hoiby, 2002). If the bacteria are liberated from the biofilm and re-investigated, they show the same sensitivity to antibiotics as the free-living forms. This resistance to antibiotics of the biofilm bacteria could be due to several factors, such as reduction of oxygen concentrations at the base of the biofilm, maybe also by providing a penetration barrier based on binding of for instance positively charged aminoglycosides to negatively charged alginate polymers; another factor may be the presence of β-lactamase from bacteria which cleaves and/or traps β-lactam antibiotics (Hoiby, 2002).

1.2.2.2. Mucoid phenotype, PMNs and oxygen radicals

A prominent feature of *P. aeruginosa* strains infecting CF patients is the conversion of non-mucoid phenotype into a mucoid, i.e., exopolysaccharide alginate-overproducing one (Cobb *et al.,* 2004). The mucoid phenotype may be advantageous for the bacteria by impeding phagocytosis and providing protection against reactive oxygen species and antibiotics. *In vivo* studies demonstrate that clearance of mucoid strains from murine lungs is diminished in comparison with non-mucoid strains, which indicates the better survival of alginate-producing strains. Alginate enhances mucin secretion by tracheal epithelial cells and may inhibit

neutrophil migration to the sites of infection (Mai *et al.,* 1993).

The conversion into the mucoid form can be caused by oxygen radicals such as those arising from H_2O_2, which are released by polymorphonuclear leukocytes (PMN) (Mathee *et al.,* 1999). H_2O_2 treatment causes a single mutation in mucA gene, which leads to production of higher levels of alginate, detectable differences in growth rate, and other features of the mucoid phenotype. This suggests that mucoid conversion is a response to oxygen radical exposure and that activation of bacterial genes by toxic oxygen radicals may serve as a defense mechanism for the bacteria (Mathee *et al.,* 1999).

1.2.3. Antibacterial response: bacterial agents, receptors and pathways

Among the products of *P. aeruginosa* known to be recognized by the receptors of epithelial cells the main role in triggering antibacterial response belongs to lipopolysaccharide (LPS), bacterial proteins pilin and flagellin. Other bacterial agents playing an important role are phenazine pigment pyocyanin and *Pseudomonas* autoinducer 1 (PAI-1). Although the other effects of these latter factors are relatively well described (O'Malley *et al.,* 2003a, b; Britigan *et al.,* 1992; Smith *et al.,* 2001), only few affected cellular pathways in eukaryotes are known and still no detailed information about triggered transcription factors is available.

1.2.3.1. Pyocyanin and autoinducer 1

The ability of *P. aeruginosa* to cause disease is partially due to its ability to produce a large repertoire of virulence factors. The production of some of these factors is regulated by quorum sensing, a cell-to-cell signaling mechanism (Passador *et al.,* 1993). In addition to delivering cytotoxic type III secretion effectors into eukaryotic cells to disrupt the host immune responses and cause cytoskeletal reorganization, *P. aeruginosa* also produces a large number of exoproducts, including proteases, hemolysin, rhamnolipids, and pyocyanin (PCN) (Lau *et al.,* 2004a and –2004b; Ran *et al.,* 2003). Two bacterial agents are of special interest in the connection with triggering pathways in epithelial cells; these are pyocyanin and autoinducer N-3-oxododecanoyl homoserine lactone (3-O-C12-HSL, also called PAI-1).

Pyocyanin (PCN) is a blue redox-active secondary metabolite that is produced by *Pseudomonas aeruginosa*. The fact of the production of some blue substance by *Pseudomonas* was noticed more than a century ago, but the role and effects of this factor are still not absolutely clear (Müller *et al.,* 1989).

It has been shown that PCN-deficient mutants cause less mortality than wild type strain. Pyocyanin's cytotoxity has been linked to its involvement in cell-mediated redox cycling

resulting in the formation of superoxide ($O_2^{-\cdot}$) and hydrogen peroxide (H_2O_2). Both NADH and NADPH are directly oxidized by PCN (O'Malley *et al.,* 2003; Britigan *et al.,* 1992). The effects of PCN are different. It inhibits α_1 protease inhibitor, which modulates serine protease activity in lung. The imbalance in protease-antiprotease activities is noticed in CF and *P. aeruginosa* patients (Britigan *et al.,* 1999). Acting as an electron carrier, PCN inhibits oxidative burst in neutrophils, which is an important defence mechanism against many microorganisms (Müller *et al.,* 1989); it also and inhibits T-lymphocyte proliferation (Mühlradt *et al.,* 1986). The other effect of PCN is inhibition of catalase activity (O'Malley *et al.,* 2003). It increases IL-8 production, but inhibits expression of the chemokine RANTES.

Although many biochemical aspects of pyocyanin activity are described, we still lack a clear picture of its effects on transcription regulation and have only vague understanding of their possible mechanisms.

The mechanism of quorum sensing enables the bacteria to regulate genes in a density-dependent manner by the production of small diffusible molecules called *Pseudomonas* autoinducers. There are two autoinducers produced by *Pseudomonas*, N-3-oxododecanoyl homoserine lactone (3-O-C12-HSL, or PAI-1), and N-butyryl L-homoserine lactone (C4-HSL, also called PAI-2). The effects of the autoinducers on the expression of the virulence factors in the bacteria are well known, but little was known about how they could affect the eukaryotic cells. It has been shown by Smith *et al*. (2001) that PAI-1 has a clear effect on the production of IL-8 in human fibroblasts and epithelial cells regulated by NF-κB. The activation of NF-κB and subsequent production of IL-8 were found to be regulated through the standart mitogen-activated protein kinase pathway. Although we do not have any information about the receptors and pathways leading to the triggering of MAPK pathway, we include PAI-1 in our scheme of interactions as an important neutrophil response stimulator.

1.2.3.2. Pilin and asialoGM1

Type IV pili, also known as fimbriae, allow *P. aeruginosa* and other pathogens to move along a solid surface by twitching motility. They extend and retract from the pole of the cell (Mattick *et al.,* 1996; Mattick, 2002; Shi, 2002). Type IV pili act as adhesins in the initial stage of binding to host cells and are one of the most indispensable components of *P. aeruginosa* colonization. Pili adhere to respiratory epithelial cells by binding to specific galactose, mannose, or sialic acid surface receptors (Wozniak and Keyser, 2004).

The type IV pili of *P. aeruginosa* are polar, polymer filaments composed of a single

protein subunit, PilA, or pilin. The distal tips of *P. aeruginosa* pili can bind to to the glycosphingolipids contained within epithelial cell membranes, thus mediating the adherence of the bacteria to epithelial cells. The glycolipids serve as carbohydrate adhesion receptors on human lung tissue (Krivan *et al.,* 1988). It has been demonstrated in *in vitro* experiments that *P. aeruginosa* binds to asialoganglioside 1 (asialoGM1) and asialoGM2, but not to the sialylated forms of GM1 and GM2 (Imundo *et al.,* 1995; Prince, 1992; Sheth *et al.,* 1994). Type IV pili recognize a disaccharide (GalNAcß1–4Gal) exposed in asialylated glycolipids such as asialoGM1. These receptors do not exist in abundance on normal epithelial surfaces, but on cell surfaces that express the mutant CF transmembrane channel regulator (CFTR) they appear in greater numbers (Imundo *et al.,* 1995). The number of asialoGM1 receptors may increase in the injured cells in patients with CF, thus reinforcing the susceptibility of cells to infection by *P. aeruginosa* expressing type IV pili.

In spite of numerous studies confirming that asialoGM1 serve as epithelial cell receptor for *P. aeruginosa*, there are data contradicting it. T. Schroeder and co-workers confirmed that asialoGM1 treatment enhanced the binding of *P. aeruginosa* strain PA103; however, no other *P. aeruginosa* strain, including eight different clinical isolates, exhibited enhanced binding to asialoGM1-treated cells (Schroeder *et al.,* 2001). This may be due to the diversity of type IV pilins in length and amino acid composition (Kus *et al.,* 2004).

1.2.3.3. Lipopolysaccharides (LPS)

Lipopolysaccharides are principal components of the outer membrane of Gram-negative bacteria. They are the dominating constituents of the outer membrane, such that about three quarters of the bacterial surface is made of LPS. LPS consists of an O-specific chain, a core oligosaccharide, and a lipid component, termed lipid A. The latter determines the endotoxic activities and essential functions for bacteria (Rietschel *et al.,* 1994). The primary structure of lipid A of various bacterial origins has been elucidated. The effects of lipid A are mediated by macrophage-derived bioactive peptides such as tumor necrosis factor alpha (TNF) or receptors like TLR4.

LPS is known as a major factor responsible for toxic manifestations of severe Gram-negative infections. LPS are the prime targets for the innate immune system. The cellular response to LPS occurs through interaction of LPS with a circulating LPS-binding protein (LBP) and a glycosylphosphatidyl inositol-linked surface receptor CD14 with subsequent activation of Toll-like receptor 4 (TLR4) (Kawai *et al.,* 2001).

1.2.3.4. Toll-like receptors and the triggered pathways. Short overview

The evolutionary ancient mechanism of innate immunity is found even in the *Drosophila* immune response, the so called Toll signaling pathway (Fig. 1). Mammalian Toll-like receptors, TLRs, homologs of *Drosophila* Toll, are key molecules for recognizing bacterial components to evoke inflammatory response (Zhang and Ghosh, 2001). TLRs are a family of pathogen recognition receptors that discriminate between diverse microbial signatures. To date, genes encoding 10 TLRs have been found in the human genome, each recognizing distinct microbial ligands (Jefferies and O'Neill, 2004). At least one ligand has been identified for each TLR, except for TLR10 (reviewed in Akira and Hemmi, 2003). During infections with *P. aeruginosa*, TLR4-binding by LPS is of major importance. All TLRs are characterized as having leucine-rich repeats extracellularly and signal via a conserved intracellular domain that they share with the interleukin-1 (IL-1) receptor family, called the Toll/IL-1 receptor (TIR) domain (Janeway and Medzhitov, 2002; Yamamoto *et al.,* 2004).

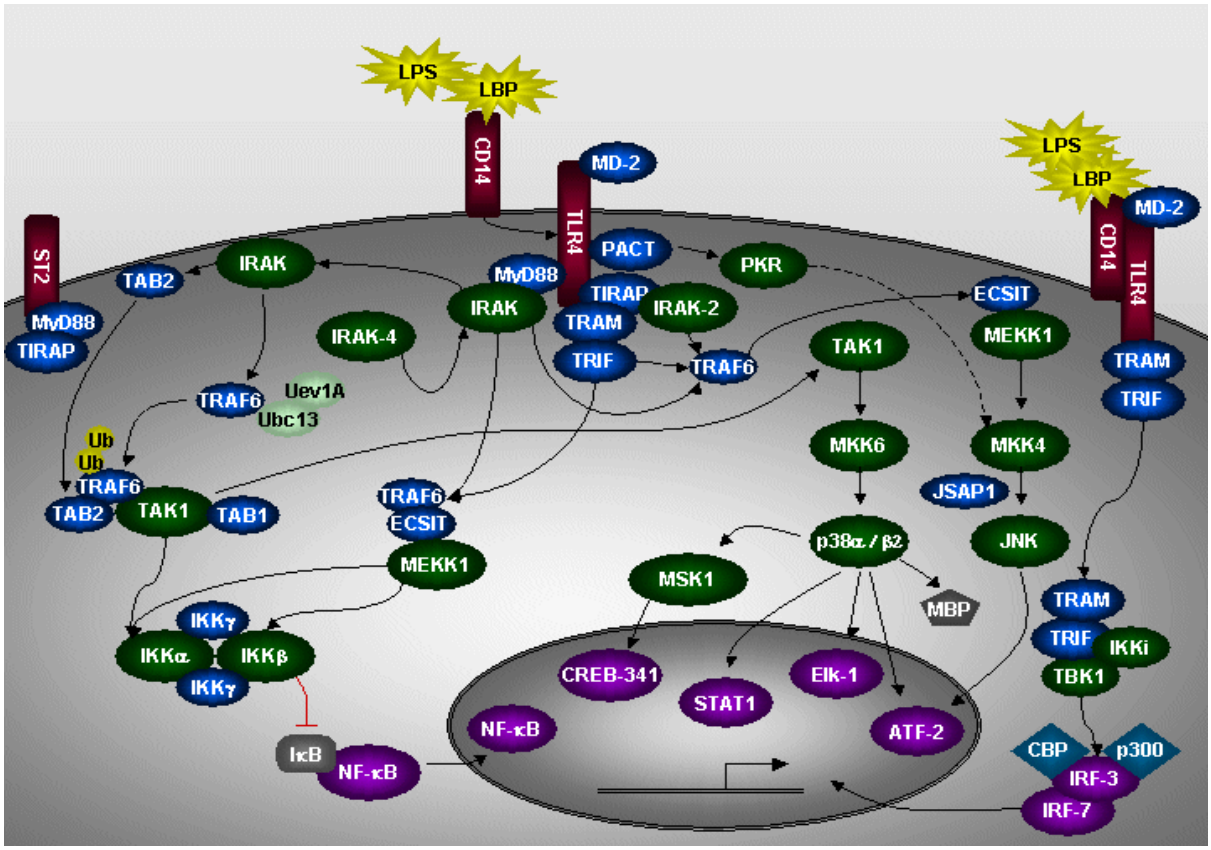


Figure 1. Scheme of interactions triggered by LPS through TLR4 (the map taken from Transpath database). Colour legend: Dark red – receptors; dark green – kinases; blue (oval) - adaptors; purple – transcription factors.

MyD88 and the two kinds of triggered pathways

There are two major pathways triggered by TLRs which are MyD88-dependent and –independent, respectively (Fig. 2). The main consequence in both cases is activation of the transcription factor NF-κB, which is crucial for the expression of pro-inflammatory cytokines and various mediators.

MyD88 is a general adaptor protein which also possesses a TIR domain. Many of the TLRs recruit MyD88 to their intracellular domains once activated, which results in subsequent recruitment of a serine-threonine kinase IL-1R-asociated kinase (IRAK-1/4). Phosphorylated IRAK dissociates from the receptor complex and associates with TNFR-associated factor (TRAF) 6 (Cao *et al.,* 1996). This leads to the activation of different pathways involving JNK/p38 mitogen-activated protein (MAP) kinase and, through the IκB kinase (IKK) complex, finally activates NF-κB. MyD88-dependent pathway is well investigated and reviewed in many publications (Yamamoto *et al.,* 2004; Palsson-McDermott and O'Neill, 2004; Zhang and Ghosh, 2001; Takeda and Akira, 2004)

MyD88-independent pathway

The experiments with MyD88-deficient cells and MyD88-deficient mice revealed the existence of MyD88-dependent and MyD88-independent pathways in LPS signaling (Kawai *et al.,* 1999). The MyD88-independent pathway regulates late NF-κB activation and the induction of genes through activation of IRF-3 transcription factor (IRF-3-dependent genes) in response to LPS.

Interferon regulatory factor-3 (IRF-3) was originally identified as a member of the IRF family that binds to the ISRE of the ISG15 (Au *et al.,* 1995; Reich *et al.,* 1987). The IRF-3 protein is ubiquitously present in a variety of tissues and phosphorylated in response to viral infection, dsRNA treatment, or DNA-damaging agents (Lin *et al.,* 1998, Navarro *et al.,* 1998). Phosphorylated IRF-3 then translocates to the nucleus, associates with the p300/CBP coactivator, and binds to the ISRE, which results in induction of several IFN-regulated genes. It has been reported that virus-induced IP-10 induction is dependent on IRF-3 and ISGF3 (Navarro and David, 1999). In the study of Kawai *et al.* (2001) it has been demonstrated that IRF-3 translocates to the nucleus in response to LPS in MyD88-deficient macrophages as well as wild-type macrophages. The IFN-regulated genes were induced by LPS in MyD88-deficient macrophages to the similar extent as in wild-type macrophages, indicating that LPS activation of IRF-3 is MyD88-independent. In contrast, induction of other genes in response to LPS was dramatically reduced (IL-6, IL-1β, and TNF-α) or abolished (COX-2) in MyD88-

deficient macrophages. Thus, LPS activates at least two signaling pathways to induce different subsets of genes; the MyD88-dependent pathway regulates expression of IL-6, IL-1β, TNF- α, and COX-2, whereas the MyD88-independent pathway regulates expression of IFN-regulated genes such as IP-10, GARG16, and IRG-1, possibly through coordinate action of IRF-3 and NF-κB (Kawai *et al.,* 2001).

The investigation of the MyD88-independent pathway revealed several additional TIR domain-containing adapters. The first was TIRAP (TIR domain-containing adaptor protein)/Mal (for MyD88-adaptor like). Unlike MyD88 which participates in all of the TLR and IL-1 receptor signaling pathways, TIRAP is an adaptor molecule specific for TLR4 signaling. Function and physiological role of TIRAP are reviewed in Yamamoto *et al.* (2004).

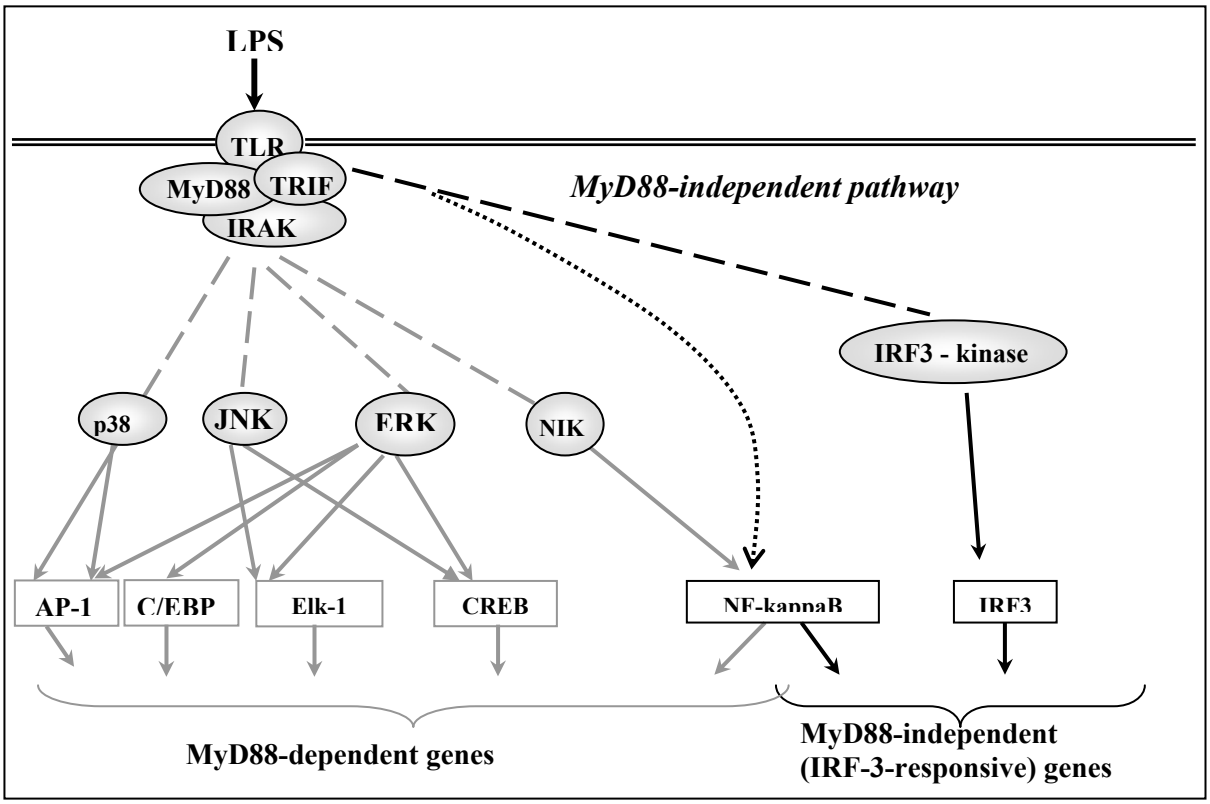


Figure 2. MyD88-dependent and –independent pathways triggered by LPS (Based on Kawai *et al.,* 2001).

The second adaptor molecule containing a TIR-domain is TRIF (TIR-containing adaptor-inducing IFN-β) (Jefferies and O'Neill, 2003). TRIF (also known as TICAM-1), interacts with TLR3 and mediates TLR3-dependent induction of IFN- β via NF-κB and IRF-3 activation (Yamamoto *et al.,* 2002; Oshiumi *et al.,* 2003). TRIF is also involved in the MyD88-independent pathway activated by TLR4. Meanwhile, a third adapter has recently been characterized, TRAM (TRIF-related adaptor molecule), which, unlike any of the other adaptors, is specific for TLR4 signaling and regulates the MyD88-independent pathway to

NF-κB and IRF-3 activation (Fitzgerald *et al.,* 2003). It also regulates late NF-κB activation in response to TLR4 (Yamamoto *et al.,* 2003). Thus, it appears that TLRs use different combinations of adaptor proteins in order to mount an appropriate response to specific microbial pathogens.

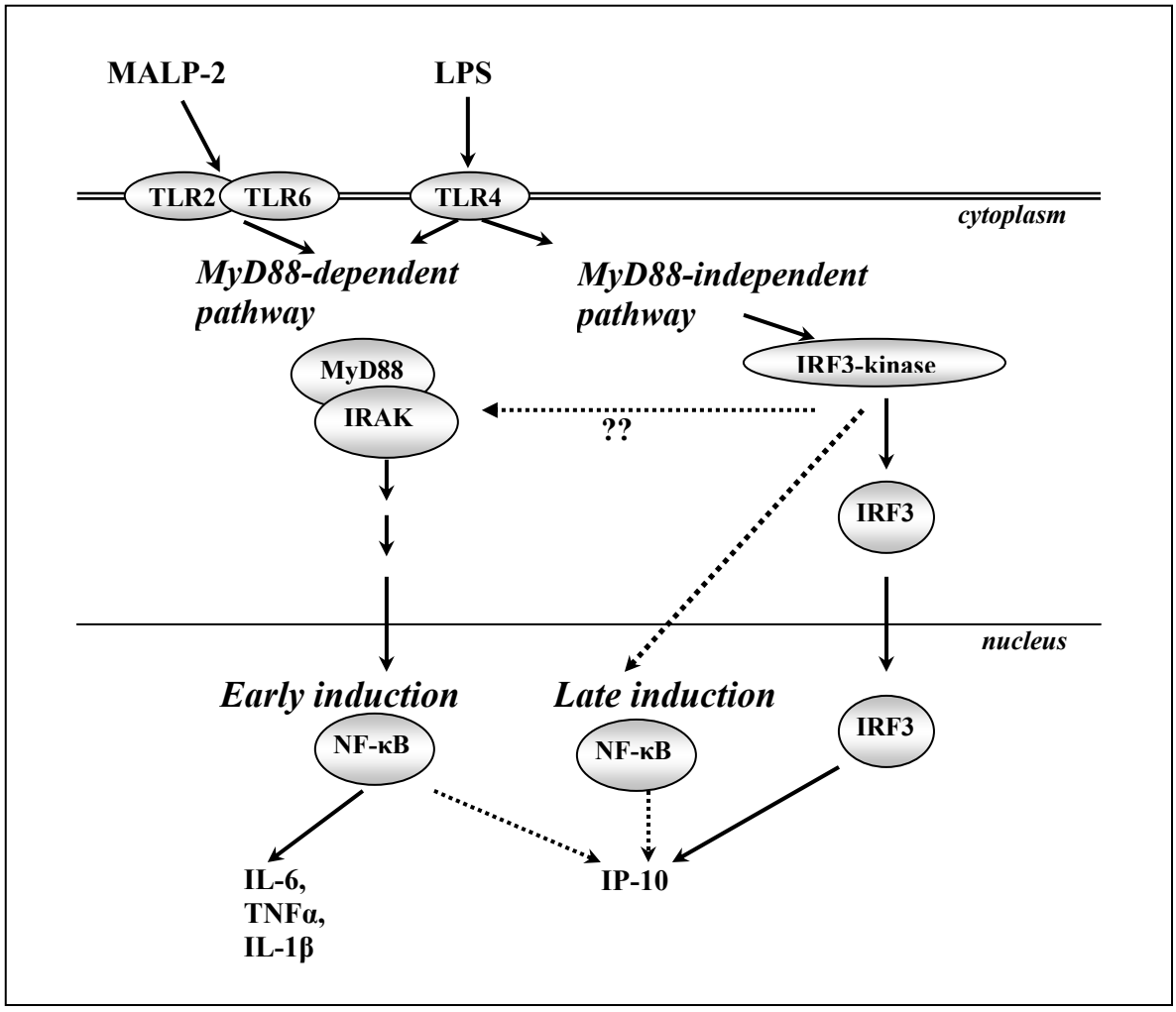


Figure 3. Model of the signaling pathways through TLR2/6 and TLR4 (adapted form Kawai *et al.,* 2001). TLR2 activates NF-κB and MAP kinases only through the MyD88-dependent pathway, whereas TLR4 activates NF-κB and MAP kinases through MyD88-dependent and -independent pathways.

MyD88 belongs not only to TLR4-triggered signaling pathways. The other members of TLR family such as TLR2 and -6 can also trigger this adapter molecule. TLR2/6 heterodimer is responsive to the activation through MALP-2 (macrophage activating lipopeptide of 2 kDa). Sato *et al.* (2000) have first demonstrated that LPS and MALP-2 act synergistically on macrophages. Kawai *et al.* (2001) suggested a possible scheme of interactions which explains the organization of MyD88-dependent and –independent pathways triggered by MALP-2 and LPS (Fig. 3)

All these observations prompt us to focus on MALP-2 in greater detail.

MALP-2 and its signaling pathways

MALP-2, macrophage activating lipopeptide of 2 kDa molecular mass, is primarily recognized by cells of the innate immune system, but also by many other cells. The recognition occurs through TLR2 and -6, two members of the family of Toll-like receptors (Sato *et al.,* 2000; www.malp-research.de). TLR 2/6 acts as heterodimer. Recognition may be facilitated by the surface molecule CD 36. MALP-2 stimulates synthesis of proinflammatory cytokines, such as IL-1, IL-6 and TNF, or chemokines, such as MIP-1 and -2, MCP-1, IL-8 and RANTES in macrophages and dendritic cells (Kaufmann *et al.,* 1999; Deiters and Mühlradt, 1999). MCP-1, e.g., is also released from MALP-2-stimulated fibroblasts. In addition, a number of surface molecules are expressed in response to MALP-2 (e.g. CD40, CD80, CD83, CD86), which are all important for cell-cell interaction, in particular with cells of the specific immune system. The effects of MALP are conducted through activation of NF-κB (Sacht *et al.,* 1998). The scheme of the signal pathway triggered by MALP-2 is presented on Fig. 4.

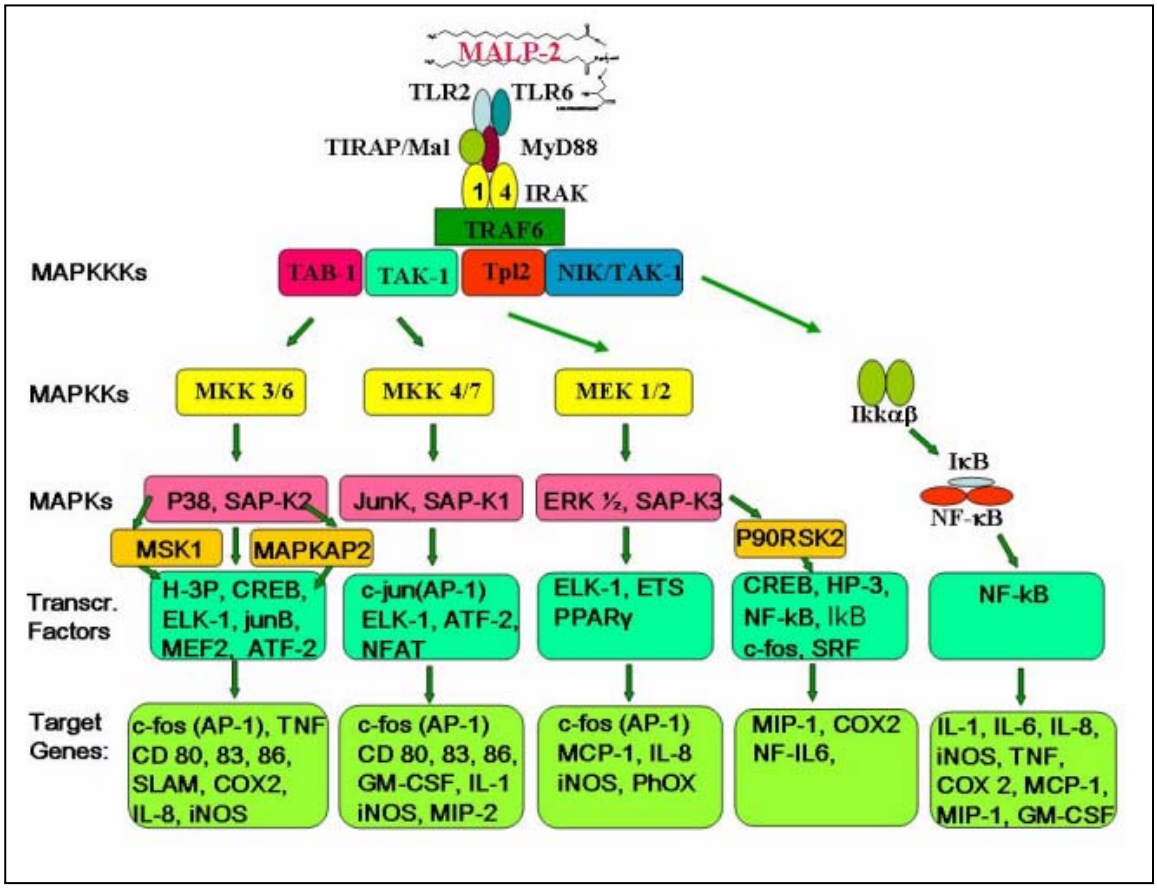


Figure 4. The scheme signaling pathways triggered by MALP-2 is taken from the website www.malp-research.de (with the kind permission of Prof. P.F. Mühlradt). Note that this scheme as well as the one for LPS triggering from TRANSFAC database is a model scheme, which presupposes, on one hand, some simplification, on the other, a more general view which summarizes the possible interactions which may happen in different cell types, organisms, etc.

1.2.4. General scheme of interactions triggered by binding of *P. aeruginosa* to human epithelial cells

While the schemes of signaling pathways triggered by LPS and MALP-2 are available in different resources (TRANSPATH® and other signal transduction databases for LPS, www.malp-research.de for MALP-2), there has been so far no scheme for the interactions triggered in the specific case of *P. aeruginosa* binding to a human epithelial cell. A full picture of such interactions is also hard to extract, if possible at all, from any databases devoted to signaling pathways like TRANSPATH®. Thus, to summarize as much presently available knowledge about this subject as possible, I suggest a general scheme of interactions caused by binding of *P. aeruginosa* to an epithelial cell (Fig. 5). This is a hand-drawn scheme based on an exhaustive literature search; I endeavored to put here all known interactions as well as the most reasonable conjectural ones. The dashed arrows represent condensed pathways where several intermediate steps were omitted from the scheme for the sake of clarity. The omitted steps are well described and can be found either as a scheme or in other representations in TRANSPATH® database.

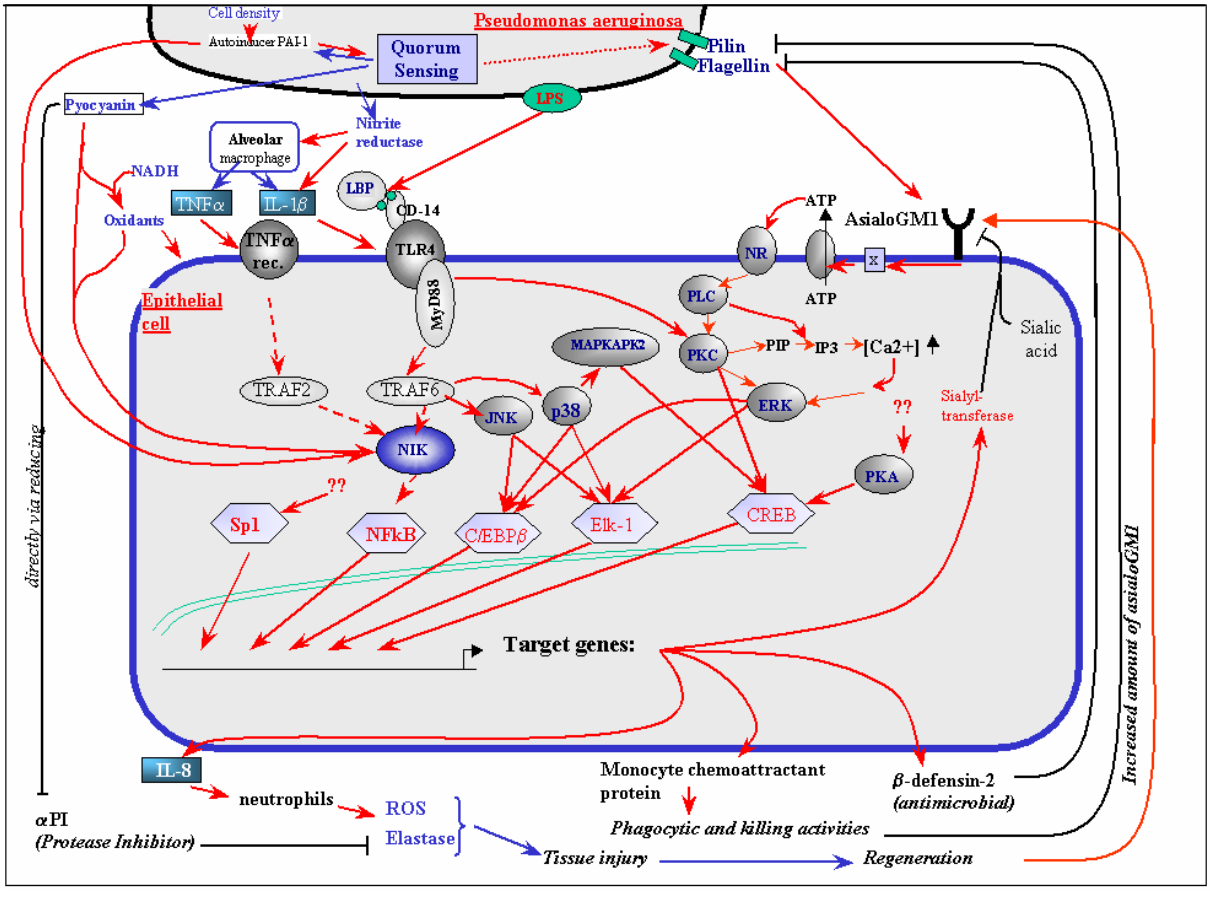


Figure 5. General scheme of interactions caused by binding of *P. aeruginosa* to an epithelial cell.

1.3. Bioinformatics: databases, methods and tools for computational approaches in biology

All bioinformatics resources can be divided into “storages” and “tools”. Storages are the depositories of biological information provided by the experimental science; these are “Databases” and “Knowledge bases”. Tools are needed to work with storages or with other experimental data.

1.3.1. Databases

Because of the relevance for work to be presented in the following, I focus on databases that store two kinds of biological data: sequence data as well as the information about signaling pathways. They will be briefly reviewed in the following.

1.3.1.1. Databases used in sequence analysis

Here we will discuss only the resources dealing with nucleic acid sequences of different functionality, as well as genes and their functions.

Nowadays there are three major nucleotide databases: EMBL (European Molecular Biology Laboratory) (Kanz *et al.,* 2005), GenBank (National Centre for Biotechnology Information) (Benson *et al.,* 2005) and DDBJ (DNA Bank of Japan) (Tateno *et al.,* 2002). They collaborate and synchronize their records on a daily basis. The rate of growth of nucleotide databases is doubling approximately every 9-12 months. Presently the databases contain information of about 100 million bases from 165000 organisms. The resource is comprehensive, but contains redundancy. A non-redundant reference sequence set for biological molecules is provided by another database, RefSeq (Reference Sequence collection, Pruitt *et al.,* 2005).

The main retrieval systems are provided by NCBI (National Center for Biotechnology Information, Wheeler *et al.,* 2005) and EBI (European Bioinformatics Institute). The NCBI Entrez is the integrated, text-based search and retrieval system for the major databases, including PubMed, DDBJ/EMBL/GenBank, RefSeq, Protein Structures, Complete Genomes, Taxonomy, and others. The EBI SRS (Sequence Retrieval System, Etzold *et al.,* 1996) is a network browser for databanks in molecular biology, which integrates and links the main nucleotide and protein databases as well as several specialized databases.

Another comprehensive resource is the database project Ensembl (Birney *et al.,* 2004). It is a source of annotation of whole genome sequences, which is providing an opportunity to browse the complete genomes of different organisms. One can not only find identified and

predicted genes there, but will be supplied also with a list of all necessary links to other relevant resources.

To get more specific information, we use more specialized databases. The specific information required for the sequence analysis of regulatory regions, considered in this work, concerns the location of regulatory sequences. There are several resources dealing with this kind of data. To list only several most important, there are: EPD (Eukaryotic Promoter Database, Perier *et al.,* 1999); DBTSS (DataBase of Transcriptional Start Sites, Suzuki *et al.,* 2002, 2004); ooTFD (Object-Oriented Transcription Factor database, Ghosh, 1998); TRRD (Transcription Regulatory Region Database, Kolchanov *et al.,* 2002); PlantCARE (Plant *cis*-acting regulatory elements, Lescot *et al.,* 2002); PLACE (Plant cis-acting regulatory DNA elements, Higo *et al.,* 1999); SCPD (*Saccharomyces cerevisiae* promoter database, Zhu and Zhang, 1999); RegulonDB (Salgado *et al.,* 2004) and PRODORIC (Munch *et al.,* 2003), transcription regulation in prokaryotes; TRANSFAC (transcription factor database, E.Wingender *et al.,* 2000). TRANSFAC presents currently the largest archive of transcription factors, their binding sites and a unique library of positional weight matrices. A database accompanying TRANSFAC is TRANSCompel, a specialized database module on composite regulatory elements (Kel-Margoulis *et al.,* 2002).

1.3.1.2. Databases on signal transduction

Signal transduction databases collect information about the transfer of signals from extracellular ligands via cell surface receptors into cytoplasm and further to different targets, mainly into the nucleus where affected transcription factors regulate their target genes (Takai-Igarashi and Kaminuma, 1999; Krull *et al.,* 2003, Lemer *et al.,* 2004). One of the earliest signaling databases was CSNDB, cell signaling networks database (Takai-Igarashi *et al.,* 1998). It was the first to come up with the idea of pathway classification based on a cell signaling ontology (Takai-Igarashi and Mizoguchi, 2004).

Presently we can list several databases as the main in the field (Table 1): aMAZE (Lemer *et al.,* 2004), CSNDB (Takai-Igarashi and Kaminuma, 1999), Reactome (previously Genome Knowledgebase) (Joshi-Tope *et al.,* 2005) and TRANSPATH (Schacherer *et al.,* 2001). Other available databases are: SPAD, Signaling PAthway Database; STCDB, a recently developed Signal Transduction Classification Database (Chen, 2004); Drastic, Database resource for analysis of signal transduction in plants (Lyon *et al.,* 2002); DOQCS, Database of Quantitative Cellular Signaling (Sivakumaran *et al.,* 2003); Dynamic Signaling Maps™, a Web-based software suite for signaling pathways, developed by Hippron™ Physiomics Inc.

The databases listed above can be called "general signal transduction databases"; one can find also several specialized databases, like Mammalian MAPK Signalling Pathways (http://kinase.uhnres.utoronto.ca/signallingmap.html), an image map depicting major components of MAPK pathways, or the Yeast Pheromone Signal Transduction Pathway (http://cbr-rbc.nrc-cnrc.gc.ca/thomaslab/sigpath.html) an image map with links to relevant references.

In the following, I would like to give short characteristics of the most important databases.

aMAZE is a WorkBench for the representation, management, annotation and analysis of information on networks of cellular processes: genetic regulation, biochemical pathways, signal transductions, biomolecular interactions, metabolic and transduction pathways.

Reactome is a curated, peer-reviewed resource of human biological processes. The basic unit of the Reactome database is a reaction; reactions are then grouped into causal chains to form pathways. The Reactome data model allows to represent many diverse processes in the human system, including the pathways of intermediary metabolism, regulatory pathways, and signal transduction, as well as high-level processes such as the cell cycle.

Table 1. General databases on signaling pathways.

Name	address	description
aMAZE	http://www.amaze.ulb.ac.be	biomolecular interactions, metabolic and signaling pathways
CSNDB	http://athos.is.s.u-tokyo.ac.jp/ace	modeling and visualization of signaling pathways
Reactome	http://www.genomeknowledge.org	metabolic and signaling pathways
TRANSPATH	http://www.biobase.de/pages /products/transpath.html	signaling pathways, network analysis
SPAD	http://www.grt.kyushu-u.ac.jp/spad/	integrated database for genetic information and signal transduction systems
STCDB	http://bibiserv.techfak.uni-bielefeld.de/stcdb/	database of information relative to the classification of signal transduction.
Drastic	http://www.drastic.org.uk/	database resource for analysis of signal transduction in plant cells
DOQCS	http://doqcs.ncbs.res.in/%7Edoqcs/	database of Quantitative Cellular Signaling; a repository of models of signaling pathways
Dynamic Signaling Maps™	http://www.hippron.com/hippron/index.html	a software suite for integration, analysis and visualization of signaling pathways.

TRANSPATH is an information system on gene-regulatory pathways, and an extension module to the TRANSFAC database system. It focuses on pathways involved in the regulation of transcription factors in different species. The states of elements of the relevant

signal transduction pathways (such as complexes, signaling molecules) are stored together with information about their interaction in a relational database.

To deal with available tools, to have an easy access to them and an overview of what is available at the moment, the Science's Signal Transduction Knowledge Environment (STKE) (http://stke.sciencemag.org/misc/about.dtl) has been proposed. The goal of Science's STKE is to identify and develop a mix of tools and approaches (algorithms, schemas, programs, and human organizational structures) that are stable, scalable, interoperable, and cost effective for providing access to information on cell signaling. STKE presents original perspectives, reviews and protocols in signal transduction research.

1.3.2. Methods and algorithms used for promoter model construction

The state of art in the search for promoter characteristics is to undertake a two-step investigation consisting of:

1. Computational search for single potential transcription factor binding sites;
2. Search for some additional characteristics (clustering of sites, combinations, distances, etc.).

The prediction of individual potential TFBS, being the first step, appears to be crucial. The used approaches are numerous and can be roughly divided into two classes:

1. "word" (consensus sequence, motif) search and
2. matrix (position weight matrix, PWM) search.

Position weight matrices are a probabilistic representation of a binding sequence. The basic idea of the identification of TFBS by any method is that the specific sites share common features and there are some consensus base pairs that almost always appear at the same position in every site (Berg and von Hippel, 1987). However, the TFBS are often highly degenerate, so one cannot apply just a consensus to evaluate the presence or absence of a specific binding pattern in a DNA sequence. Thus, for the prediction of binding sites the position weight matrix (PWM) technique has been developed (Stormo *et al.,* 1982; Harr *et al.,* 1983; Staden, 1984; Stormo and Hartzell, 1989; Goodrich *et al.,* 1990; Hertz *et al.,* 1990; Stormo, 1990). The comparison of PWM and motif consensus approaches in the search for TFBS demonstrated the advantages of PWM (Stormo *et al.,* 1982; Stormo, 2000). There are several approaches to building PWMs. Basically most of them are similar to the approach suggested by Staden in 1984 (Staden, 1984). The PWM is extracted from a set of aligned TFBS; one counts the number of occurrences of each base in each position and builds a base frequency table. The number of rows is 4 (as the number of bases), and the number of

columns is equal to the length of the TFBS. Based on the frequencies, we can estimate the probabilities of each base occurring at each position in true binding sites. In the most frequently used approaches, this estimate is a log-probability, i.e., natural logarithm of the value from the frequency table normalized by the number of sequences in the TFBS set. The main problems of the PWM are that (i) they depend on the initial set of TFBS, which may as well contain some false positives (since experimental data has its rate of false results); and (ii) it is hard to find an optimal cut-off value for PWM application. Methods for the optimization of the cut-off value for a given PWM were suggested by P. Bucher (1990) and by Pickert *et al.* (1998). Tsunoda and Takagi (1999) further refined Bucher's method and calculated the optimal cut-off values for the 205 matrices for vertebrate TFs from the TRANSFAC database.

PWMs have been used for the prediction of the binding affinity for numerous bacterial (Stormo, 1990) and eukaryotic TFs (Fickett, 1996; Frech *et al.,*1997; Tronche *et al.,* 1997; Tavazoie *et al.,*1999). Presently PWMs are routinely used in resources like TRANSFAC database and in some of the accompanying software (http://www.gene-regulation.com/pub/programs.html). Despite the obvious advantages of PWMs, the majority of existing PWMs provide a low level of both sensitivity and specificity (Frech *et al.,* 1997).

The basic assumption of the PWM approach is the statistical independence of the base pairs in the TFBS; the validity of this assumption is doubtful, in the view of recent results that demonstrate the existence of dependencies between the positions (Ben-Gal *et al.,* 2005; Bulyk *et al.,* 2002). This is a significant drawback of this type of model; the other negative characteristic is the relatively low number of parameters that can be used; this leads to high numbers of false positive predictions (the models appear to be under-fitted).

The positional dependence in a TFBS sequence is taken into account by other approaches: fixed-order models, such as Markov models, hidden Markov models (Ohler *et al.,* 1999;Ohler and Niemann, 2001; Hughes, 2000; Salzberg *et al.,* 1998,1999) and variable order models (variable Markov models and variable order Bayesian networks) (Ben-Gal *et al.,* 2005). Surprisingly, PWM perform comparable and often even better results than fixed-order Markov models based on the valid inner-dependency assumption (Ben-Gal *et al.,* 2005). The explanation is in the over-fitting of the fixed-order Markov models due to their large dimensionality, given the limited amount of training data (Ben-Gal *et al.,* 2005).

Thus, there are two major problems in the prediction of individual TFBS: (i) taking into account the inner dependency of positions within the TFBS and (ii) finding the optimal number of parameters allowing to find a balance between under- and over-fitting of the models.

An interim solution is suggested by the variable order models, which take into account the statistical dependencies, but only those, that are found to be statistically significant, thus avoiding the strong over-fitting. The variable order Markov model (VOM) was originally suggested by Rissanen (1983) for data compression and was later adapted for predictions and identifications, modeling genetic texts including TFBS and protein coding regions (Buhlman and Wyner, 1999, Orlov and Potapov, 2000, Orlov *et al.,* 2002). The variable order Bayesian network (VOBN) model was suggested by Ben-Gal and coauthors (Ben-Gal *et al.,* 2005) as a generalization of VOM. In both VOM and VOBN the order may vary from position to position based on their context (i.e., the specific nucleotides observed in the vicinity), which is the main difference from the fixed-order Markov models where the order does not depend on the position or context. VOBN model can be also considered as generalization of PWMs, fixed-order Markov models and Bayesian networks, as all these models appear to be special cases of the VOBN under certain conditions. The model suggested by Ben-Gal and coauthors not only takes into account the dependencies between the positions, but also allows to decrease the number of parameters to overcome the problem of over-fitting. In spite of these fine achievements, the suggested model appears to be only slightly better than PWM approach: in the example shown by the authors (identification of $\sigma 70$ binding sites in *Escherichia coli*) the difference between the true positive rates for VOBN and PWM models is about 3% (while the rates themselves are around 44-47%). Thus, neither of the methods provides really high sensitivity and it is questionable whether one can insist on the definite superiority of the new (VOBN) method. In spite of the fact that the difference is significant, there remains a question: are we really interested in difference of 3% on the background of 45%? Thus, despite the great theoretical interest of the approach of VOBN, PWMs still do not lose their superior role in the TFBS prediction. Similarly, the HMM-based approaches as well as those based on neural networks (ANN), which were believed to make a revolution in motif finding about 5-10 years ago, did not provide significant benefits in comparison with the PWM search.

The algorithms for recognizing regulatory modules (i.e., combinations of TFBS) (also referred in some papers as regulatory motifs) can be roughly divided into two categories:

- generative approaches (unsupervised learning; modeling sequence characteristics analyzing a positive training set of regulatory sequences) and
- discriminative approaches (supervised learning; modeling the difference between regulatory and non-regulatory sequences).

The difference is in the requirement for training sets. Generative approaches have a definite drawback in comparison with discriminative ones requiring larger data sets for analysis; moreover, they may tend to "classify everything", thus giving rise to a large number of false positives. On the other hand, the discriminative techniques require as an input not only a positive training set, but also a negative one, i.e. a set of sequences that are known to be non-regulatory. This kind of negative information is normally difficult to obtain, because the experiments are usually designed in a way that allows to confirm functionality rather than to disapprove it. Negative results *per se* are usually not published. The problem of negative training sets is, therefore, of crucial importance and is a disadvantage of the discriminative approaches. Given the mentioned shortcomings of the both approaches, the final decision which of them to apply depends mostly on the quality and quantity of available data sets.

The other way of classification the approaches is connected with the way of defining parameters. In this sense, we can consider again two groups of techniques:

- approaches where the parameters have to be predefined by a user (e.g., sliding window approaches, which require from a user specification of several parameters such as the width of the sliding window; the threshold for consideration a matching motif as a genuine, the minimum (or maximum) number of occurrences expected to appear in the investigated sequence(s);
- self-learning algorithms, when the parameters are extracted directly from the dataset (Markov models, hidden Markov models, Bayesian techniques).

It is a general opinion (shared by informaticians and most of bioinformaticians), that the less user-dependent is the searching machine, the better it is, so the necessity of *a priori* predefined parameters is a certain disadvantage of the approaches of the first group. The approaches of the second group, such as hidden Markov models (HMMs), are able to learn the parameters directly from the training set in a user-independent manner, which definitely looks like a benefit. But any kind of “machine learning” makes high demands to the training sets: to make use of this learning ability one has to provide a good (in terms of both quality and quantity) positive training set. Unfortunately, nowadays the relative scarcity of knowledge about exact location of regulatory sequences and of experimentally proven binding sites makes this a serious problem. Consequently, at the present moment the approaches based on HMMs require that the user specify the trade-off parameters rather than allowing the parameters to be learned from data (Bailey and Noble, 2003; Frith, 2001, 2002). Thus, as we can see, it is difficult to make full use of superiority of self-learning approaches. That may only mean that the time of the complete machine learning is still to come. (It remains

questionable whether it is in general possible, but this question is beyond the scope of this work).

Giving a short overview of the methods developed so far, we should mention the methods based on: (i) distinct oligonucleotide distribution (van Helden *et al.,* 1998); (ii) differential distribution of individual known TFBS and TATA boxes (Kel *et al.,* 1995, Kondrakhin *et al.,* 1995, Prestridge, 1995); (iii) Bayesian statistics (Crowley *et al.,* 1997, Qin *et al.,* 2003); (iv) Neural networks (Lukashin *et al.,* 1989, O'Neill, 1991, Matis 1996). There are numerous papers devoted to the regulatory module recognition (Bailey and Noble, 2003, Brazma *et al.,* 1998, Fickett and Wasserman, 2000, van Helden, 2003, van Helden *et al.,* 1998, Klingenhoff *et al.,* 2002, Krivan and Wasserman, 2001, Wagner, 1999, Werner *et al.,* 2003), as well as a list of developed tools (see Tables 2, 3). In spite of that, we still lack a standard method which would enable us to produce promoter models. This may indicate that the existing approaches have their distinct shortcomings and that, thus, the field is still open for new ideas.

1.3.3. Tools for promoter modeling

The range of tools applied to promoter modeling corresponds to the steps of this process. As it has been mentioned in the beginning of the previous section, we have to search first for single potential transcription factor binding sites and then search for some additional characteristics (clusters of sites, distances, etc.).

1.3.3.1. Tools for motif and TFBS search

For the prediction of potential TFBS we can apply either the tools searching for TFBS with the help of PWMs, or searching for motifs. In the first case, we not only get information about the location of predicted sites, but assign the corresponding factors; in the second case, after having identified motifs, we have to apply some additional tools to identify which factors may bind to these motifs. Thus, the tools applying PWMs are not only more effective than the motifs searching tools (see the previous section), but appear to be more straightforward and comfortable.

There are many approaches developed for automatic discovery of motifs. Among the best and most popular are: GIBBS Sampler (Thompson *et al.,* 2003, 2004; Lawrence *et al.,* 1993); MEME (Bailey and Elkan, 1994); Consensus (Hertz and Stormo, 1995); Motif sampler (Thijs, 2001) (Table 2,A). The iterative Gibbs sampling algorithm was suggested by Lawrence et al. (1993). It became one of the most popular and frequently used algorithms for motif search. The corresponding tool can find several distinct patterns simultaneously, and if all the input parameters are selected correctly, the results are very reliable. Unfortunately, the tool requires

a prior knowledge of the expected number of motifs, the number of occurrences of the motifs in each sequence, as well as the number of motifs in each subset of sequences. All these parameters are sometimes hard to estimate, if possible at all, and the result depends on them dramatically.

The other popular approach was proposed by Hertz and Stormo (1995) and was implemented in a tool called Consensus. The method is based on calculation of weight matrix for common pattern under sampling the best L-mers.

Another interesting tool is WordSpy (Wang and Zhang, 2005) which is intended to discover all over-represented (degenerate) words in a large set of sequences (biological or English language); it identifies discriminative words with negative sequence data, selects biological meaningful DNA motifs using gene expression data and evaluates DNA motifs with genome-scale random sampling analysis.

MatInspector (Quandt *et al.,* 1995), TFBIND (Tsunoda *et al.,* 1999), AliBaba2 (Grabe, 2002), PROMO (Messeguer *et al.,* 2002), and Match (Kel *et al.,* 2003) head the list of tools searching for TFBS (Table 2, B). MatInspector and Match are based on the same methodology, but the advantage of Match is that it uses the whole library of TRANSFAC PWMs and allows to construct own PWMs. AliBaba was constructed as a tool for constructing specific matrices for each analyzed sequence; it starts from a dataset of known binding sites and ends with the identification of a potential new binding site. Thus, it applies PWMs, but first constructs them. AliBaba is based on the collection of binding sites from TRANSFAC; unfortunately, only a rather old version of the database is used and, to our knowledge, the tool is not actively maintained anymore. Following a similar strategy, the recently published P-Match combines scanning with PWMs with sequence string comparison (Chekmenev *et al.,* 2005).

The other approaches (different from PWM search) to TFBS identification are also applied: e.g., TFScan (Table 2) does not apply PWMs, but rather looks for TFBS by taking a sequence and the name of one taxonomic group and performs a fast match of the TRANSFAC sequences against the input sequence, optionally allowing mismatches.

Some of the above-listed approaches base their predictions on pure statistical basis (e.g., Ann-spec, POCO, POBO, AliBaba, Improbizer, BioProspector); the others use the evidence of comparative genomics, which makes their predictions more biologically reliable (e.g., ConSite, FOOTER, CompareProspector).

Table 2. Tools searching for motifs and TFBS

name	address	description
A. Searching for motifs		
Ann-spec	http://www.cbs.dtu.dk/services/DNAarray/ann-spec.php	locates common DNA regulatory patterns in a set of promoter region sequences
Bio-Prospector	http://ai.stanford.edu/~xsliu/BioProspector	Discovering DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes
Compare-Prospector	http://ai.stanford.edu/~iliu/CompareProspector/	BioProspector which incorporates comparative genomics features to be used for higher eukaryotes
Consensus	http://bioweb.pasteur.fr/seqanal/interfaces/consensus-simple.html	identification of consensus patterns in unaligned DNA and protein sequences
FOOTER	http://biodev.hgen.pitt.edu/cgi-bin/Footer/Footer.cgi	finds mammalian DNA regulatory regions using phylogenetic footprinting
FUZZNUC	http://bioweb.pasteur.fr/seqanal/interfaces/fuzznuc.html	nucleic acid pattern search
GIBBS Sampler	http://bayesweb.wadsworth.org/gibbs/gibbs.html	allows to identify motifs in DNA or protein sequences
Improbizer	http://www.cse.ucsc.edu/~kent/improbizer/improbizer.html	motifs in DNA or RNA sequences that occur with improbable frequency
MELINA	http://melina.hgc.jp/	motif extraction from promoter regions of potentially co-regulated genes
MEME	http://meme.sdsc.edu/meme/website/intro.html	tool for discovering motifs in groups of related DNA
Motif Sampler	http://www.esat.kuleuven.ac.be/~thijs/Work/MotifSampler.html	find over-represented motifs in the upstream region of a set of co-regulated genes
Motif-Search	http://compbio.cs.princeton.edu/bindsites/webform/index.html	comparative analysis of methods for representing and searching for transcription factor binding sites
MotifViz	http://biowulf.bu.edu/MotifViz/	an analysis and visualization tool for motif discovery : three motif discovery programs, Clover, Rover and Motifish
WordSpy	http://cic.cs.wustl.edu/wordspy/	identifying transcription factor binding motifs by building a dictionary and learning a grammar
B.Searching for TFBS		
AliBaba2	http://www.alibaba2.com/	prediction of transcription factor binding sites by context dependent matrices generated from TRANSFAC
CONREAL	http://conreal.niob.knaw.nl.	identification and visualization of conserved transcription factor binding sites
ConSite	http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/	prediction of regulatory elements using cross-species comparison
Mat-Inspector	http://www.genomatix.de/products/MatInspector	Searches for TFBS using PWMs
Match™	http://www.biobase.de/cgi-bin/biobase/transfac/start.cgi	Searches for TFBS using PWMs
MDscan	http://ai.stanford.edu/~xsliu/MDscan/	finds protein-DNA interaction sites from ChIP-on-chip targets
PAINT	http://www.dbi.tju.edu/dbi/tools/paint/	Analyzes the TF-binding site occurences for over/under-representation compared to a reference
Patch™	http://www.gene-regulation.de/ cgi-bin/pub/programs/pmatch/bin/patch.cgi.	Searches for TFBS using PWMs
P-Match™	http://www.gene-regulation.de/ cgi-bin/pub/programs/pmatch/bin/p-match.cgi.	Searches for TFBS using PWMs
POBO	http://ekhidna.biocenter.helsinki.fi:9801/pobo	transcription factor binding site verification with bootstrapping
POCO	http://ekhidna.biocenter.helsinki.fi/poco	Can be applied to one or two clusters of promoters of co-regulated genes to detect nucleotide patterns (motifs) that are distinguishly represented in them.
PROMO	http://alggen.lsi.upc.es/recerca/promo/intro-promo.html	identification of putative transcription factor binding sites
TFBIND	http://tfbind.ims.u-tokyo.ac.jp/	searching transcription factor binding sites using PWMs.
TFExplorer	http://tfexplorer.org/	1. shows predicted TFBS in the promoter regions, along with their phylogenetic footprinting information. 2. searches for genes that have a given sequence pattern in their promoter regions using the motif-searching method.
TFScan	http://web.umassmed.edu/cgi-bin/biobin/tfscan	locates potential transcription factor binding sites

1.3.3.2. Tools for further promoter analysis

The number of tools for the identification of complex regulatory patterns (i.e., clusters or other combinations of motifs/binding sites) is rather limited (Table 3).

BioProspector (Liu *et al.,* 2001), Co-Bind (Guha Thakurta and Stormo, 2001), MITRA (Eskin and Pevzner, 2002) have been developed for the prediction of statistically over-represented pair-wise combinations. The first two are based on an extension of Gibbs sampling techniques, looking for motifs with some flexible distance between them. The MITRA approach is based on enumeration of l-mers and building a mismatch tree data structure to split the space of all possible patterns into subspaces. All these tools were successfully applied to the prediction of regulatory elements in prokaryotes and yeast; it remains questionable whether they can be applied to the higher eukaryotes.

Table 3. Tools for analysis of TFBS combinations.

name	address	description
CisMols Analyzer	http://cismols.cchmc.org/peak-web/index.jsp	identification of compositionally similar cis-element clusters in ortholog conserved regions of coordinately expressed genes
Cister	http://zlab.bu.edu/~mfrith/cister.shtml	Cis-element Cluster Finder
Comet	http://zlab.bu.edu/~mfrith/comet	Cluster Of Motifs E-value Tool. It finds statistically significant clusters of motifs in a DNA sequence
Co-Bind	http://ural.wustl.edu/~dg/co-bind.html	program for identification of target sites for cooperatively binding transcription factors
MITRA	http://www1.cs.columbia.edu/compbio/mitra	Mismatch tree algorithm for finding regulator elements in DNA sequences.
PredictRegulon	http://210.212.212.6/prindex.html	prediction of the regulatory protein binding sites and operons in prokaryote genomes
BioProspector	http://ai.stanford.edu/~xsliu/BioProspector	Prediction of regulatory motifs from co-regulataed genes in prokaryotes or lower eukaryotes
CompareProspector	http://ai.stanford.edu/~iliu/CompareProspector	an extension to BioProspector which incorporates comparative genomics features to be used for higher eukaryotes
FrameWorker	http://www.genomatix.de/products/GEMSLauncher/index.htmlhttp://www.genomatix.de/products/GEMSLauncher/index.html)	a software tool enabling the detection of common TFBSs in conserved arrangement (frameworks) from a set of DNA sequences, solely based on sequence analysis.

The other tools deal with the prediction of clusters of motifs/TFBS. CisMols Analyser (Jegga *et al.,* 2005) allows to identify *cis*-regulatory modules, called CisMols, that occur in groups of coexpressed or related genes within evolutionarily conserved *cis*-regulatory regions.

The tool Cister (*Cis*-element Cluster Finder) is a HMM-based and searches for hetero-clusters of TFBS (Frith *et al.,* 2001). Comet (Cluster Of Motifs E-value Tool) is written by

the same authors (Frith *et al.,* 2002), also aims to detect clusters of *cis*-elements, thus these two programs are quite similar. The most important differences are that the output of Comet is easier to interpret and that Comet indicates the statistical significance of its predictions using an E-value. On the other hand, Cister integrates all possible arrangements of cis-elements in the cluster, whereas Comet just considers the most probable arrangement.

MSCAN (Alkema *et al.,* 2004) is developed for measuring the statistical significance of non-overlapping TFBS combinations in a window. Evolutionary conservation is not taken into account.

From all tools listed above, only CisMols Analyser takes into account not only statistical over-representation, but evolutionary conservation of the sequence elements; this is a serious omission of the other programs, because involvement of comparative genomics is now the state of the art in the identification of regulatory sequences.

2. RESULTS

This chapter describes the developed approaches to promoter model construction and their application to the investigation of several defensive systems of eukaryotic cells. Since the connection of all parts may be not self-evident, I precede some sections with the explanation of motivation.

2.1. Subtractive approach to positional weight matrix generation

2.1.1. Motivation

To generate a positional weight matrix one needs to align the sequences of known binding sites for the transcription factor. The better the binding sites are aligned, the better will be the matrix. However, some TFs are characterized by extremely degenerate binding patterns; sometimes it is principally impossible to align the whole set of the available binding sites. Binding patterns may be weak for different reasons, one of them may be that the training sets for deriving the patterns are too large and heterogeneous, representing different subgroups of binding factors and/or classes of binding sites. For instance, different heterodimers of C/EBP isoforms with each other or with other bZIP transcription factors may exhibit distinct sequence specificity (Shuman *et al.,* 1997). If we could separate the subgroups of the whole heterogeneous training set in advance, it would be possible to construct matrices for each such subgroup, each matrix being of higher quality than a matrix for the "mixed" set. The first approach is to use the biological knowledge about the different isoforms of the TFs. But the biological evidence does not always help to separate the mentioned subgroups: the patterns for different isoforms, which can be expected to exhibit different binding patterns, do not necessarily show it (for instance, the binding sites for several isoforms of C/EBPα, -β, -γ and -δ, do not form discrete groups). Nevertheless, we can often see very similar, if not identical, patterns present in the large part of the training set(s), even if not assigned to any isoform subclass. This means that there can be a way of computational differentiation between the subclasses based on the pure sequence analysis. Thus, there is a need for an approach which would be able to define the subgroups of the binding patterns in the whole training set and derive matrices for each such subgroup. In the end, we will have to use a set of matrices instead of one matrix to characterize a binding site of one TF, but, as we demonstrate below, this leads to an increase of both sensitivity and specificity of the search.

2.1.2. Description of the approach

2.1.2.1. Subtractive approach to matrix generation

The positive training set consisted of sequences containing experimentally proven transcription factor binding sites for the investigated transcription factor taken from the TRANSFAC® database (see *Methods*). The whole set has been put into GIBBS Motif Sampler for DNA (http://bayesweb.wadsworth.org/gibbs/gibbs.html), searching for motifs of 10 nucleotides length. The length of the motifs has been selected according to the previous information about the "standard" TFBS which tend to have the consensus sequence of the length of 10bp (this can be changed for certain types of the TFs if there is information about other preferable length of the binding sites). The sequences with the motifs were aligned and put to the "Matrix generation" subroutine of Match™ (Kel *et al.,* 2001). The obtained matrix has been used for the search in the training set with high cut-offs (the cut-offs should be estimated for each type of matrix and should provide re-identification of 50-80% of true positives; one can use the minFP cut-off suggested by the "Matrix generator"). All the sequences found by the new matrix were subtracted from the whole set, and the remaining set was used again for the motif searching. The procedure was repeated until 90% of all true

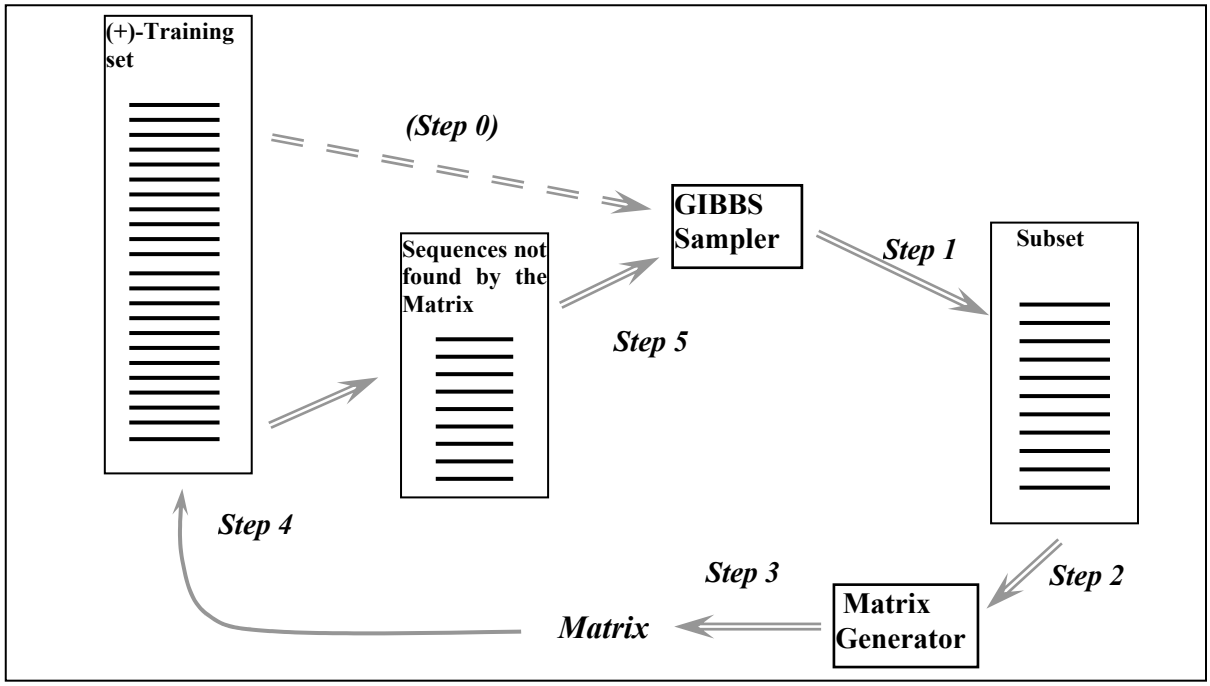


Figure 6. The scheme of the subtractive approach. *Step0.:* The whole training set is searched by Gibbs Sampler for the motifs of the length 10. *Step 1.* The sequences with the found motif are collected in one subset. *Step 2.* The sequences of the subset with the found motif are aligned and put into the "Matrix Generatror" of the Match tool. *Step 3.* The matrix is generated. *Step 4.* The obtained matrix is used for the search in the training set with high cut-offs providing re-identification of 80% of true positives. *Step 5.* The sequences not found by the matrix form a new inquiry set which is put to the Gibbs Sampler. From this point the cycle is repeated until 90% of all sequences of the initial training set are utilized.

positive sites in the set were used for constructing a matrix (Fig.6).

When working with positional weight matrices it is important to identify a threshold with which a matrix will search for a certain rate of true positives (or false negatives). The approach described below has been developed for the identification of thresholds for single matrices as well as for sets of matrices. In the latter case, any set of matrices is the set of PWMs for the same TF (for example, obtained by the subtractive approach), which we want to use simultaneously.

2.1.2.2. Defining thresholds for a set of PWMs.

For the search with PWMs we used the tool Match (Kel *et al*., 2001). The threshold for each matrix of the set has been set to a very low level (0.6 or less for both the core and the matrix similarity). This allows to get all possible potential hits in all sequences. From the Match™ output we derived a list of hits for matrix similarity thresholds for each sequence in the training set, the name of the matrix being assigned to every hit (Fig. 7). For each sequence of this list only the hits with the highest scores were taken for the further analysis.

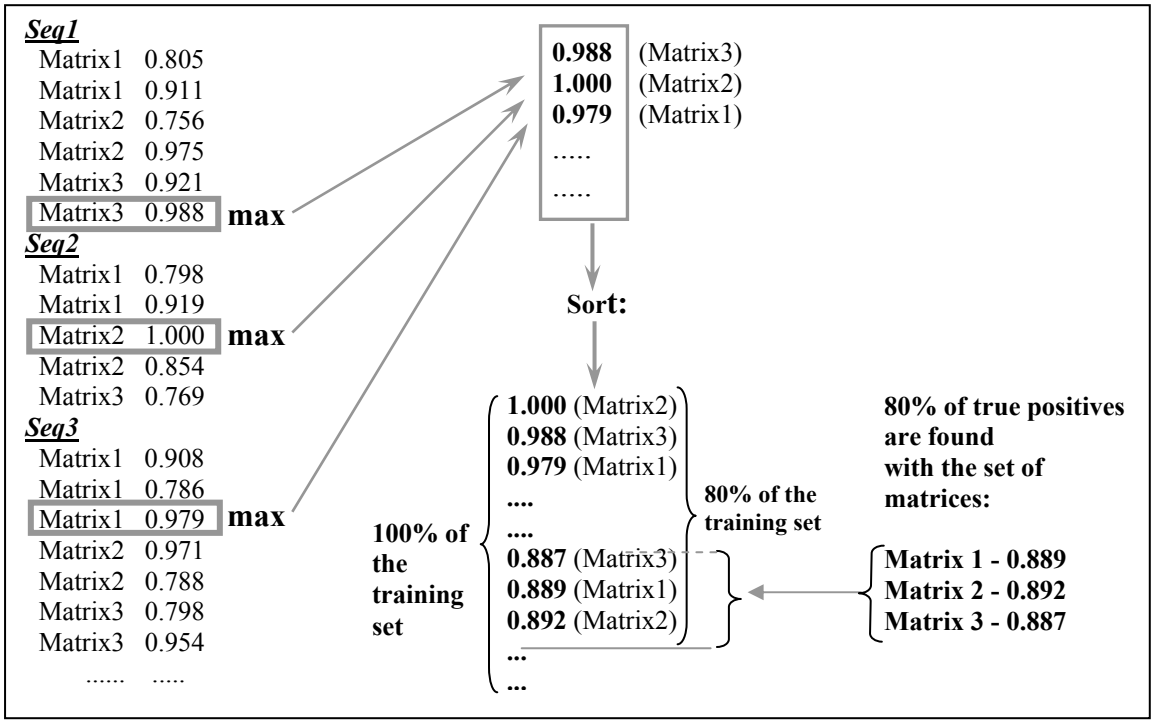


Figure 7. Defining thresholds for re-identification of the selected rate of true positives for a set of PWMs. The PWM set consists of three matrices: Matrix1, 2 and 3. After the search for predicted TFBS in the training set (containing the true positive sequences) with the Match tool using very low thresholds for every matrix we obtain a list of hits for matrix similarity thresholds for each sequence, the name of the matrix being assigned to every hit (first column). For each sequence of this list, only the hits with the highest scores are found (in the red frames); they form a new list (second column, in the red frame). This list of the highest scores is sorted and is ready for the definition of the thresholds for the desired level of TP (here 80%). Here, Matrix 2 with the score 0.892 appears on the level of 80% of TP; for the other two matrices, the scores of their next occurrences upwards are taken.

This list of the highest thresholds with which every sequence of the training set has been found was now sorted by decreasing score and ready for the definition of the thresholds for the desired level of TP.

2.1.3. Application to C/EBP matrix re-evaluation

We applied the subtractive approach to re-evaluation and improvement of the existing matrices for C/EBP binding sites. We had a special interest in this transcription factor for several reasons: first, C/EBP is known to participate in immune response; second, it is one of the target factors in the pathways triggered by the increase of $[Ca^{2+}]$ which mediates the asialo-GM-dependent response to *P. aeruginosa* invasion; and at last, it is known to be a part of the composite element with NF-κB. Thus, we were interested in a better quality of the predictions of the C/EBP binding sites than it was possible with standard matrices from TRANSFAC.

C/EBP is known to have several isoforms (C/EBPα, -β, -γ and -δ), but the sequences of their binding sites do not form any discrete groups. They cannot be aligned in order to derive a matrix that would be specific for any isoform (for instance, only C/EBPα sites), nor any combination of them (which could be useful in the case of C/EBPγ and -δ for which only very few binding sites are known). The whole set of the binding sites is too heterogeneous and the matrices existing in TRANSFAC reflect more the history of the binding sites accumulation in the database than the real situation.

When this work was done first in 2002 (Shelest & Wingender, 2003), there were 164 binding site entries for C/EBP in TRANSFAC (release 6.1) in non-artificial (genomic) sequences and 8 matrices for them, but all of them exhibited rather weak consensi. We decided to make more precise matrices for C/EBP dividing the whole set into subgroups using the subtractive approach.

When I applied the approach outlined above, it revealed a set of 4 matrices, each of them searched for a subset of C/EBP sites. Comparison of the consensi derived from these matrices shows that they represent distinct sequence patterns (Table 4). The set of available TFBS-containing sequences contained 154 items (see "Supplementary materials/subtractive approach/Subtractive_seq_sets.doc", S1).

As it has been mentioned above, C/EBP often appears as a constituent of the C/EBP-NF-κB composite element; as NF-κB is a crucial element of the innate immunity pathways, we were particularly interested in identifying C/EBP sites within the composite elements. Thus, we undertook an attempt to construct a specific matrix for them. We took only the sites

contained in the composite elements of NF-κB-C/EBP type as they were documented in the database TRANSCompel 6.1 (Kel-Margoulis *et al*., 2002; see "Supplementary materials", "Table_comp.doc" for the sequences) and constructed a specific matrix for this subset ("CEBP_comp"). Interestingly, the consensus sequence of this matrix perfectly corresponded to the consensus of the matrix derived by the first round of subtractive approach ("cebp_new") (see the matrices in *Appendix 1a*).

Because of the significant increase in C/EBP training sequences in TRANSFAC and because of refinement of tools used (Match and Gibbs Sampler) since then, it was decided to revisit this part of the study. Thus, we re-made the work with the new set of input sequences, corresponding to the 220 TFBS in non-artificial sequences in the TRANSFAC version 9.1 (spring 2005). The whole new set consisted of 193 sequences containing C/EBP binding sites (some of the TRANSFAC entries did not have a link to a sequence or did not contain a sequence), each site being prolonged by 10 nucleotides to each side (see "Supplementary materials/subtractive approach/Subtractive_seq_sets.doc", S2). The first round of motif search selected on the first step two practically coinciding subsets with 68 and 72 motifs, correspondingly, but both sets gave weak matrices; to define the matrix more precisely, I put the 68-motif set again to Gibbs Sampler and got a subset of 62 sequences containing same motif (see "Supplementary materials/subtractive approach/Gibbs_sub10-1.txt",and "--/Gibbs_sub10–2.txt"); the corresponding matrix ("cebp_sub10", see *Appendix 1a*) matched with 107 sequences of the positive training set. The next rounds gave motif subsets of subsequently 36 (matrix "cebp_sub11" re-identifying 93 sequences, overlapping in small part with the first subset), 23 (matrix "cebp_sub12" re-identifying 55 seq), and 8 sequences (matrix "cebp_sub_13" re-identifying 9 seq) (all Match^{TM} outputs can be found in "Supplementary materials/subtractive approach/match_subXX.txt"). All 4 matrices, when applied together, covered 177 sequences of the true positive set, leaving 16 sequences as a rest, in which hardly any motif could be found (see *Appendix 1b,* table "Re-identification of C/EBP TFBS by the 4 new matrices").

The consensus sequences derived from the different "old" and "new" matrices are compared in Table 4. It should be noticed that the newly created matrix "cebp_sub10" corresponds to the matrix derived from C/EBP/NF-κB composite elements CEBP_comp; both are similar to the matrix Cebp_new. Both matrices Cebp_new (2002) and cebp_sub10 (2005) were created from the first round of the subtractive approach (thus, representing the largest subset of sequences). Matrix "cebp_alternative" corresponds to the matrix "cebp_sub11". Interestingly, the matrix "cebp_rest10" which was derived in the 4th round of the subtraction in 2002 can be

nicely aligned now with the matrix "cebp_sub12" (2005) which was derived in the 3rd round; vice versa, the matrix "cebp_rest1" (3rd round in 2002) now corresponds to the matrix "cebp_sub13" (4th round in 2005). We will return to the comparison of consensus sequences and matrices in the *Discussion*.

Table 4. Comparison of the consensus sequences for C/EBP binding sites derived from positional weight matrices (created in 2002 and 2005). The 3 pairs of consensi were aligned; the fourth does not fit to any of the other.

pair number	Round of sub-traction	year	name	consensus
	-	2002	CEBP_comp	S H N V N R T T G C A C A A
1	1	2002	Cebp_new	T T G T G Y A A
		2005	Cebp_sub10	T T R C M C M A
2	2	2002	cebp_alternative	C A T T K C S Y C A K N
		2005	Cebp_sub11	A T T K C Y T M A K
3	4 3	2002	cebp_rest10	G G S D G A G G W G
		2005	Cebp_sub12	D G C A S A G G
4	3 4	2002	cebp_rest1	T A T T K G C T
		2005	Cebp_sub13	W N T G A T T G C T

Sets of matrices for comprehensive search

To make the search for binding sites more comprehensive we combined the matrices in such a way that for a defined rate of true positives (TP) a minimal rate of false positives (FP) is achieved and the overlap between the subsets recognized by individual matrices is minimized (Fig. 8). FP is represented here by f_r, the frequency of matches per nucleotide in the control set of random sequences. To estimate the error, we conducted all measurements in 10 sets of random sequences, 50 sequences in each set. We retrieved the f_r for all single C/EBP matrices available in TRANSFAC as well as for the newly created using the subtractive approach (for the set of matrices created in 2005 for the new set of input sequences) and for the sets of 2 (sub10+sub11), 3 (sub10+sub11+sub12) and 4 (sub10+sub11+sub12+sub13) matrices. The results are presented in Table 5.

As can be seen, the use of the combination of 4 matrices reduces the FP rate by about two thirds compared to any of the previously used (M00109-100621) and newly created individual search patterns.

The approach has been picked up in the context of the Master thesis of D. Fredrich. He newly developed and implemented an own algorithm replacing GIBBS Sampler that is suitable for large-scale use. The resulting tool can be applied to the construction or/and improvement of matrices for any set of binding sites. Applying this program to C/EBP sites,

he came up with a set of very similar motifs as those shown here, proving the robustness of the approach.

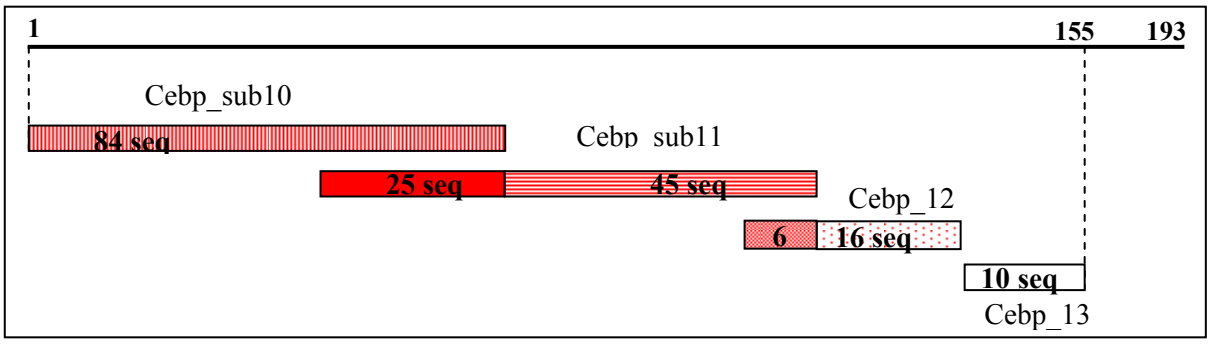


Figure 8. Re-identification of 80% of true positives by the set of 4 matrices obtained by the subtractive approach with the distribution of hits for every matrix in the set.

Table 5. Frequencies of sites per nucleotide found in control sets of random sequences, for the re-identification of 80% of true positives (i.e., FN rate allowed = 20%).

	f_r (x10^{-3})													
	M00109	M00116	M00117	M00159	M00190	M00201	M00621	sub10	sub11	sub12	sub13	set2	set3	set4
Random1	19	17	19	20	20	17	16	18	52	36	32	15	11	5
Random2	17	15	18	19	19	15	17	18	49	36	31	15	10	5
Random3	17	14	16	20	20	15	15	17	48	37	31	14	10	5
Random4	18	15	17	20	20	16	17	18	47	36	30	14	10	6
Random5	18	16	18	18	20	17	16	17	48	35	31	15	10	6
Random6	17	16	18	19	19	16	16	17	48	35	31	15	10	5
Random7	17	15	17	21	20	16	18	19	48	39	31	14	9	5
Random8	17	16	18	20	20	17	18	19	49	36	29	15	10	6
Random9	17	14	16	21	19	15	16	17	47	36	29	13	11	5
Rand_10	18	14	17	19	20	17	16	18	47	34	28	14	11	5
Mean	17,5	15	17,4	19,7	19,7	16,1	16,5	17,8	48,3	36	30,3	14,4	10,2	5,3
stdev	0,7	1	0,97	0,95	1	0,9	0,5	0,8	1,5	1,3	1,3	0,7	0,63	0,48

2.2. Distance distributions

2.2.1. Motivation

Model construction is based on the consideration of combinations of transcription factor binding sites (TFBS).The problem to find a functional combination of TFBS *in silico* is not trivial because of the unreliability in the prediction of individual TFBS. As a consequence, the main problem remains the high number of false positive predictions, independently of whether we use a motif- or a positional weight matrix search. Thus, we encounter a situation

when the real characteristic under consideration (a TFBS, in this case) may be detected by a tool, but cannot be recognized on the background of noise. This problem is easier to solve when we have not only the sequence characteristics of the motifs themselves, but can add some other independent constraints, for instance the distance between distinct occurrences of the same or different motifs.

For distinguishing the signal from the noise we need to model the level of noise. In sequence analysis one can either go for computer simulations using random (or randomized) sequences, or analytically calculate the result for the case of random distribution. The latter has an obvious advantage of a comprehensive approach, whereas the randomization approaches are normally heuristic; on the other hand, it is not always possible to describe analytically the investigated processes. Here we show that it is possible to model analytically the distribution of distances in random case. Such a "random distribution" will represent the noise.

2.2.2. Calculation of theoretical distance distribution

We investigated the dependency of the number of TFBS pairs on the distance at which they occur (distance distribution). For any set of sequences, which we consider, there will be some background noise of the distance distribution due to unavoidable false positives among the predictions for TFBS. The main purpose of this part of work was to see whether it is possible to differentiate the real signal (distances of real TFBS pairs) from the noise signal (distances of false positive TFBS pairs).

We suppose that the coordinates of the false positive TFBS are random. Let us consider a model system with random distribution of points on a segment, where the points are representing the sites thus neglecting the extension of TFBS. We consider a sequence of the length L, in which we find M_A sites of type A and M_B sites of type B, A and B being distributed randomly. The pairs can be represented as dots of intersection in a square with the side L, M_A dots being distributed along the one side and M_B along the other. It is evident, that we then get $M_A M_B$ pairs, the maximal number of positional combinations being L^2 (supposing that different sites can occupy the same place). We want to estimate what will be the distribution of the distances between sites A and B. In other words, we want to know how many intersections in average can occur at some distance *d.* In our graphical representation this means the number of pairs (dots of intersection of M_A and M_B) which occur on a line corresponding to the distance d (green dots in Fig. 9a). The number of pairs found at the distance 0 (f_0) will be equal to L (Fig. 9a, red dots) and the number of pairs at the distance d,

where $1 \le d \le L$-1, will be equal to 2(L-d), as we consider the pairs A-B and B-A as the same.

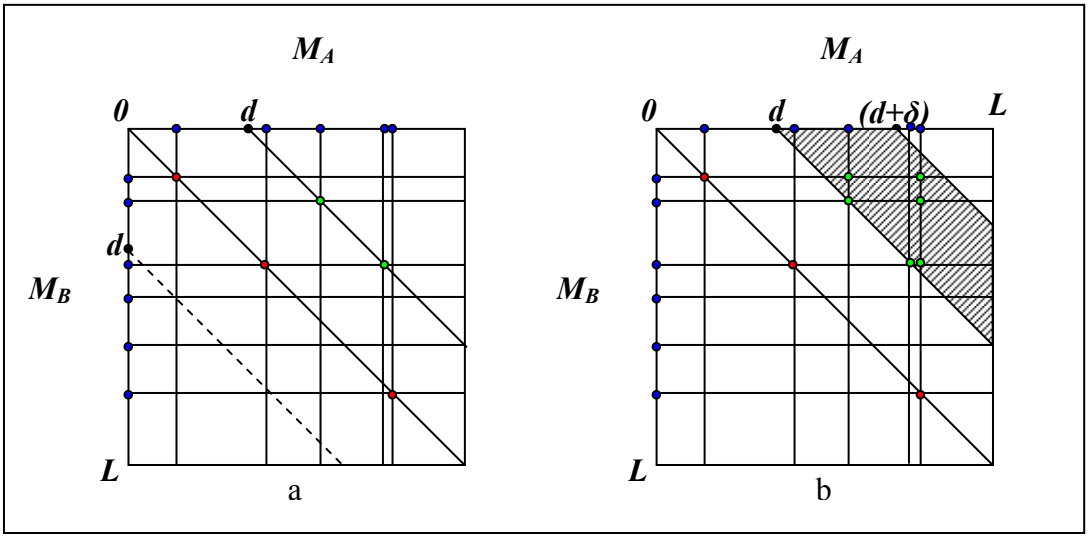


Figure 9. Graphical representation of the pairs' occurrence. M_A and M_B are the numbers of the sites of the types A and B, correspondingly. The sites of the both types are distributed randomly along the sequence with the length L; the blue dots represent their occurrences. The intersections of the lines corresponding to the coordinates of every site represent the occurrences of the pairs of the A-B type; the red dots show the cases when A and B coincide (d=0). d is a distance between the two sites of a pair.

a. Consideration of pairs strictly on the distance d.
b. Consideration of pairs in some distance interval, from d *to* $d+\delta$.

Let f_d be an average number of pairs at distance d. Note that the number of pairs found at the distance 0 f_0 will be equal to L, and the number of pairs at the distance d, where $1 \le d \le L$-1, will be equal to 2(L-d). Therefore

$$\begin{cases} f_d = 2M_A M_B (L-d)/L^2, \ 1 \le d \le L-1 \\ f_0 = M_A M_B / L \end{cases} \qquad (1)$$

The distances in genuine TFBS pairs (composite elements) are not rigid. To allow slight shift of the sites, we consider the pairs on some distance interval, δ (Fig. 9b). Let $f_{d,\delta}$ be the average number of pairs on the distance interval from d to $d+\delta$. Analogously to (1), it is easy to show that

$$\begin{cases} f_{d,\delta} = 2(\delta+1)\left(L-d-\dfrac{\delta}{2}\right)\dfrac{M_A M_B}{L^2}, \ 1 \le d \le L-1 \\ f_{0,\delta} = \left(2(\delta+1)\left(L-\dfrac{\delta}{2}\right)-L\right)\dfrac{M_A M_B}{L^2} \end{cases} \qquad (2)$$

Note that factor $M_A M_B / L^2$ is the product of frequencies of TFBS A and B (M_A / L and M_B / L, respectively) which were predicted in one sequence. In the case of set of N sequences

this factor must be changed to $\sum_{i=1}^{N} M_{A,i} M_{B,i} \Big/ NL^2$. To estimate the error of the predictions, we undertook computer simulations that showed that in the case when $M_A << L$ and $M_B << L$ the distribution of $f_{d,\delta}$ is close to normal and has the standard deviation $\sigma \approx 1.3\sqrt{f_{d,\delta}/N}$ (see *Appendix 2*).

The theoretically calculated distribution of distances between random sites will be called further on ***random distance distribution***.

2.2.3. Comparison of random distance distributions with the distance distributions in the control set of random sequences

We checked the quality of our theoretical predictions comparing the calculated results with direct measurements in a control set (see *Methods*). For each pair concerned, we measured the number of occurrences of distances at which the pair has been found in the set. The results of the comparison are shown in Fig. 10.

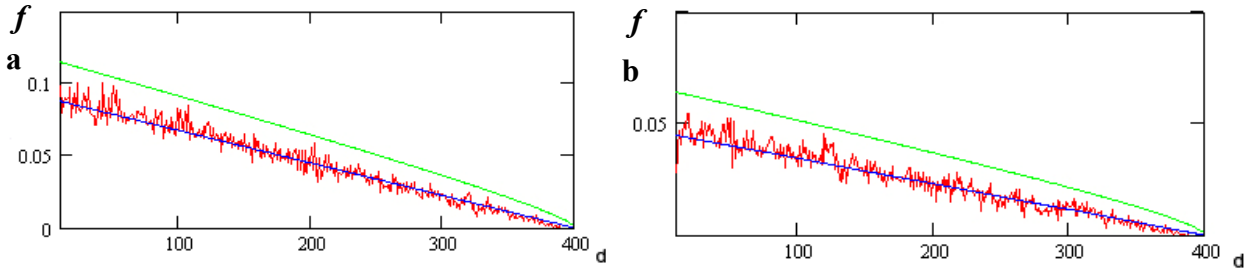


Figure 10. Comparison of the measured (red line) and calculated (blue line) TFBS distance distributions for two examples: (a) AP-1–ETS; (b) AP-1–NF-κB. The green line indicates the calculated distribution plus 3 standard deviations.

2.2.4. Application of the distance distribution approach

In order to understand the behavior of the TFBS pairs and whether it is possible to use the distances as a distinguishing feature we have investigated the behavior of real, experimentally proven TFBS pairs which have been shown to bind cooperating transcription factors (TF) - composite elements. Composite elements (CE) are combinations of two or more transcription factor binding sites which provide synergistic action of the TFs, qualitatively different from a purely additive effect. The most comprehensive collection of composite elements can be found in the TRANSCompel database (Kel-Margoulis *et al.,* 2002). We compare the distance distributions in real composite elements with the theoretically calculated distributions of distances in random cases.

2.2.4.1. Distance distributions in composite elements

The advantage of the real (experimentally proven) composite elements (CE) is that we know the positions of genuine binding sites for cooperating TFs. This is the best model of true positives for promoter model construction. The main disadvantage of real CE (at the present stage) is that for most of them the number of known instances is very low (usually less than 5). Only 8 sets retrieved from TRANSCompel are relatively large (7-19 hits in the database). Thus, we have analyzed the following 8 composite element types: AP-1–ETS, AP-1–NFAT, AP-1–NF-κB, NF-κB–C/EBP, NF-κB–IRF, IRF–PU.1, NF-κB–Stat and NF-κB–HMG I(Y).

In the reported composite elements we know the positions of the binding sites, but when we search for them using the positional weight matrix approach we always find some additional predictions, which may be false positives. Actually, we do not know whether they are false because of the imperfection of the predictions, or whether they are true, which could be a feature of a promoter: it can be saturated with the sites of the needed type(s), for the TFs to have additional sites to bind (see *Discussion* for more details). The aim of our investigation was to see whether the distribution of distances in the sequences with reported CEs is different from random and what are the features of this distribution.

Sets of CE-containing sequences are considered separately for each composite element. We searched for potential TFBS with a PWM approach as described in *Methods* and measure the distances for all found TFBS pairs of the corresponding type. The difference from the random distribution by more than 3σ is considered as over-representation. The results of the comparison for the considered CEs can be seen in Fig. 11. The peaks exceeding the 3σ-line are considered in more details in Table 6.

2.2.4.2. Coincidence of the dominating peaks and the true positive distances

Each of the true positive sequences contains 1-2 experimentally proven pair(s) of TFBS, which we consider as the true positive examples. We will call them further on “true positive (TP) pairs”. They should not be mixed up with the other pairs found in the same sequences. Analogously, we will use the term “TP distance” to describe the distance in a TP pair.

All TP distances for each composite element are listed in Table 7. These TP distances can be merged to peaks as long as they occur close to each other. We merged them if the difference was lower than the length of one site (10bp) (Table 7). This has been made for the sake of easier comparison of the obtained data.

Table 6. Re-identification of true positive distances in the over-represented peaks.

Composite element	True positive peaks	Found over-represented peaks	% of re-identified TP pairs	Number of additional over-represented peaks
AP-1-ETS	**5-19** 32 **52-57** 92	**6-19** **51-63** 107-120 144-161 206-219	86%	3
AP-1-NFAT	**7-11**	**5-18** 125-146	100	1
AP-1-NF-κB	**9-18** **30-36** **47**	**5-61**	100	0
	127	**127-163**		
IRF-PU.1	**5-7**	**2-8**		0
NF-κB- C/EBP	**7-21** **34** 60	**0-40**	86	0
	72-88 **109-111**	**74-135**		
NF-κB-HMG I(Y)	**0-2** **23**	**0-14** **22-35** 190-210	100	1
NF-κB-Stat	**4-6** **25** **53** **73** **109**	**0-8** **17-29** **46-57** **74-86** **105-117** 254-263	100	1
NF-κB–IRF	**11-28** **48** **60** **75**	**7-35** **48-81**	100	0

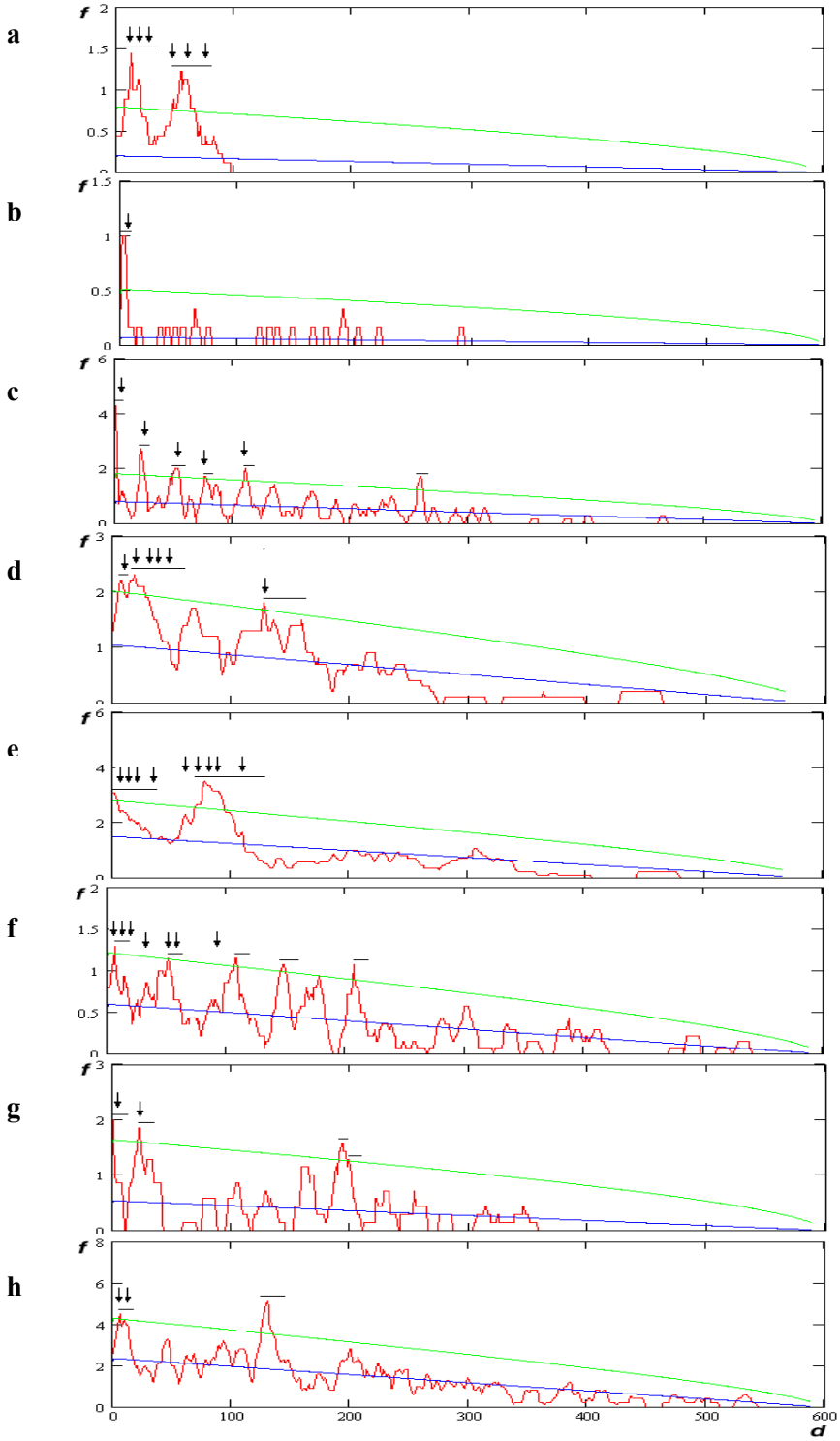


Figure 11. The results of measurement of the average number of pairs per sequence *f* ($f_{d,\delta}$) normalized by the number of sequences) at distance *d* in the CE-containing sets (red line) are compared with the results of calculation for these sets (blue line) plus 3 standard deviations (green line). For the over-represented distances, the peaks are shown with δ (black line above the peaks) (for instance, if the peak is found at the distance 50 (which is marked with the red line) with δ=25, any distance in the interval 50-75 is considered as over-represented; the black line shows the actual size of the over-represented peaks). The arrows mark the TP distances. a.NF-κB–IRF; b. IRF–PU.1; c. NF-κB–Stat; d. AP-1–NF-κB; e. NF-κB–C/EBP; f. AP-1–ETS; g. NF-κB–HMG I(Y); h. AP-1–NFAT.

In the plots of Fig. 11 one can see that most of the TP distances coincide with the peaks of over-represented distances. δ was selected for each composite element pair according to the biological evidence, i.e. taking into account the distance shifts occurring in real composite elements. The rate of re-identification is rather high: from 86 to 100% of all TP distances have been re-identified in the over-represented peaks (Table 6).

Table 7. List of true positive (TP) distances and peaks.

Composite element	TP Distances	TP Peaks
AP-1–ETS	5,6,7,8,11,13,15-17,19,32,52,53,57,92	5-19,32,52-57,92
AP-1–NFAT	7,7,7,7,8,8,9,9,9,9,10,10,10,10,11	7-11
AP-1–NF-κB	9,10,12,18, 30,34,36,47, 127	9-18,30-36,47,127
IRF–PU.1	5,5,6,6,7,7,7	5-7
NF-κB–C/EBP	7,12,13,15, 21,34,60,72, 78,83,87,88, 109,111	7-21,34,60,72-88, 109-111
NF-κB–HMG I(Y)	0,1,1,1,1,1,2,23	0-2,23
NF-κB–Stat	4,5,6,25,53,73,109	4-6,25,53,73,109
NF-κB–IRF	11,11,15,18,20,24,28,47,60,75	11-28,48,60,75

2.2.4.3. Potentially false predictions

For some types of CEs (NF-κB–IRF, AP-1–NF-κB, IRF–PU.1 and NF-κB–C/EBP), all peaks of over-represented distances coincided with the TP peaks. For the other composite elements, we observed additional peaks of over-represented distances (Table 6), which could be interpreted as false positive predictions. Interestingly, they occurred only on longer distances. We will return to this observation in the *Discussion*.

2.3. Other anti-false-positive measures

The main problem of promoter model construction is the numerous false positives. Developing our approaches, we applied some anti-false-positives measures, in addition to the described approaches to improve the specificity of single-site detection (2.1) and to consider distance correlations (2.2):

- identification of "seed" sequences
- phylogenetic conservation
- subclassification into complementary sequence sets.

In the following, we will comment on each item in more detail.

2.3.1. "Seed" sequences

Initially the idea of "seed" sequences was exploited because of the desire to make use of preexisting biological knowledge about the genes used in the positive training set and also because of doubts in the reliability of the available data set.

The basic idea of the promoter model construction is that the co-regulated genes should share some common sequence pattern in their regulatory regions. Thus, the ideal positive training set should represent the genes, which are co-regulated. Unfortunately, we rarely have information about the actual co-regulation and have to substitute it with the information about the co-expression of genes. This is the first problem. The second problem is that different experimental approaches differ in their reliability. The microarray analysis is not absolutely reliable (Pritchard *et al.,* 2001; Lee *et al.,* 2000; Pan, 2002; Draghici *et al.,* 2003), so we can expect that not all of the reported genes may be relevant for the considered system. Thus, collecting and using the information for construction of a positive training set, we have to make two assumptions: (i) that the co-expressed genes represent co-regulated genes; (ii) that the experimental data is reliable. Both statements are doubtful, but we do not have other sourses of information. On the other hand, some genes are already known to be relevant according to additional published evidence. This allows us to judge about their reliability with more confidence. The most reliable genes can be used as a "seed" set, i.e. a subset of sequences where the occurrence of the features characterizing the co-regulation are expected with the highest probability (we took it as 100%).

Fig. 12 illustrates how the approach of the "seed" set is applied as a filtering technique to the promoter model construction: first, we identify all TFBS pairs that are present in all sequences of this "seed" group (see *Methods*) (Fig. 12, step 2). Further on, we search for the found pairs in the whole (+)-training set (Fig. 12, step 3). In the next step we make a search in the (-)-training set for those pairs that were found in at least 80% of the (+)-training set (Fig. 12, step 4), choosing only those which showed the lowest percentages in the (-)-training set (Fig. 12, step 6).

Using this approach, we could avoid being drowned by a flood of pairs, most of which would be of minor importance. For example, in the case of the antibacterial response of human epithelial cells to *P. aeruginosa* the initial number of pairs in different intervals which could be found in the whole (+)-training set was nearly 37,000; this huge number was reduced by the application of the "seed" approach by at least two orders of magnitude: depending on the "seed" the number of considered pairs varied from 50 to 400. In the next steps, this number was reduced by another order of magnitude (Table 8).

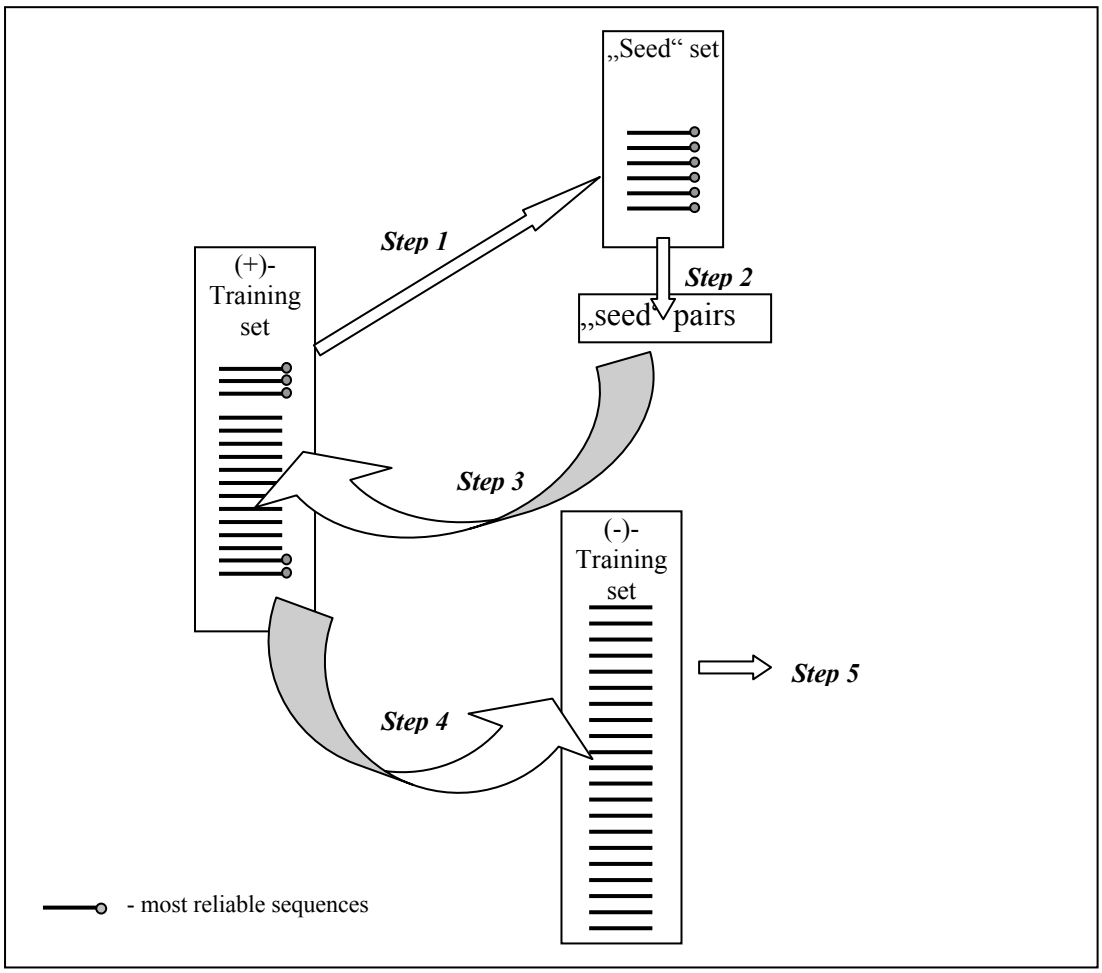


Figure 12. Algorithm of the search for common pairs using seed sets. Step 1. Selection of a “seed” set. Step 2. Identification of all pairs in the “seed” set; only those, which are found in 100% of the “seed” sequences, are taken into further consideration. Step 3. Search for the selected pairs in the whole (+)-training set. Only those pairs, which are found in more than 80% of sequences of the (+)-training set, are taken into further consideration. Step 4. Search for the “survived” pairs in the negative training set. Only those which are present in less than 40% of sequences are left. Step 5. The list of the common pairs is ready for the next analysis.

Table 8. Stepwise filtering of pairs.

Pairs found on different steps of the search	No of found pairs
Pairs found in the whole training set in all distance intervals	~ 37000
Pairs found in the “seed” set in all distance intervals (step 2 on the fig.12)	~ 400
“Seed” pairs in more than 80% of the training set (step 4 on the fig.12)	~180
“Seed” pairs in more than 80% of the training set and less than 40% of the negative training set (step 6 on the fig.12)	4

Each "seed" is characterized by its own set of pairs. To ensure the robustness of the obtained results, we undertook the "leave-one-out" test, removing consecutively one sequence of the "seed" set (for the combined "seed" sets which included human and mouse orthologs we excluded simultaneously both orthologous sequences). This has been repeated for each sequence (or ortholog pair). Only the robust pairs have been taken into further consideration.

This general approach was applied first when we tried to construct the promoter model for genes that are involved in the epithelial response on a bacterial infection (*P. aeruginosa*). It is described in more detail in the corresponding paragraph (2.5.1.1., Selection of the "seed" set).

2.3.2. Complementary pairs

To facilitate the search for combinations we tried to exploit the concept that subsets of phenomenologically co-regulated promoters may be subject to differential regulation. If the response of the cell is mediated through at least two distinct pathways, it is logical to suppose that there are subsets of promoters activated by each of them. The subsets may not be obvious from the expression data or from any other observation, but in some cases (as in the one analyzed here, where we have two different pathways triggering the same response; see Fig.5) one can presuppose the existence of two or more subsets, each of them possessing an own combination of TFBS. These combinations will be complementary in the sense of their occurrence in the set (Fig.13).

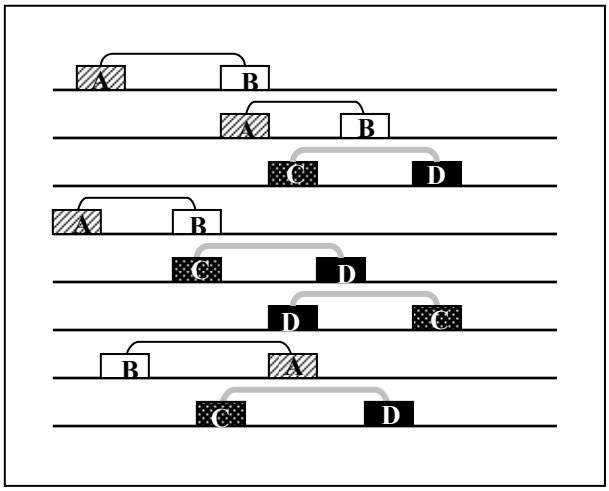


Figure 13. Complementary pairs. A, B, C and D are transcription factor binding sites, which form two sorts of pairs (A-B and C-D). These pairs are complementary in the sense of occurring in complementary subsets of the whole set.

For simplicity, we considered only binary combinations (pairs) of TFBS, but the search for combinations of higher order would be likewise possible and make the model more

specific. Moreover, detection of complementary pairs enables to identify corresponding complementary subsets of sequences, thus to shed light on some features of the ascending regulatory network.

2.3.3. Phylogenetic conservation

Evolutionary conservation of a (potential) TFBS is generally accepted as an additional criterion for a predicted site to be functional (Levy *et al.,* 2001; Hardison, 2003; Pennacchio and Rubin, 2003). However, some recent analysis of the human genome reported by Levy and Hannenhalli (2001,2002) and our own observations made for short promoter regions have shown that only about 50% (Levy *et al.,* 2001), 64 % (Hannenhalli and Levy, 2002) or 70 % (Sauer *et al.,* in preparation) of the experimentally proven binding sites are conserved. Missing between 30 and 50 % of all true positives may seem to be acceptable when analyzing single TFBS, but if one constituent of a relevant combination of TFBS belongs to a non-conserved region, we will loose the whole combination from all further analyses.

The observed fact is that functional features are not necessarily bound to conserved regions, as long as we speak about primary sequence conservation. Dealing with such degenerate objects as TF binding sites, one should not expect an absolute conservation of their binding sequences. From the functional point of view, it seems to be more reasonable to expect that not the sequences, but the mere occurrence of binding sites and/or their combinations as well as (perhaps) their spatial arrangement would be preserved among evolutionarily related genomes. That is the approach that we use in the present work, completely refraining from sequence alignments. We search for those pairs of TFBS, which can be found in human and corresponding mouse orthologous promoter regions, considering the promoter as a metastring of TFBS. We took a feature (the pair of TFBS) into account only if we could identify it in both orthologous promoters, not taking into consideration in what region of the promoter it appeared; we also did not try to align metastrings of TFBS symbols, since they may be interrupted by many additional predicted TFBS (no matter whether they are true or false positives). While this work was in progress, we found a very similar approach in the work of Eisen and coworkers (Chiang *et al.,* 2003; Moses *et al.,* 2003), who searched for conserved “word templates” in the transcription control regions of yeast. We believe that switching from primary sequence preservation to the conservation of higher-order features like clusters of TFBS is the next step in development of the approaches of comparative genomics.

2.4. Promoter model construction

2.4.1. Identification of pairs with defined mutual orientation

We consider all possible pair-wise combinations of TFBS in each sequence, as described in *Methods*. Multiple occurrences of a pair in a sequence are not taken into account. The distances between the constituents (r_1, $r2$) may be optimized by this approach or be selected in advance using the approach of the distance distributions. Since we did not model the mutual orientations analytically, we use the control ((-) - training) set for comparison.

Let us consider two TFBS *m* and *n* located in a distance range from r_1 to r_2 (where $r_1 \le r_2$) on either strand of DNA (+ or -). We can denote the sets of sequences containing pairs in different relative orientation as $A_{m^+,n^+}(r_1,r_2)$, $A_{m^+,n^-}(r_1,r_2)$, $A_{m^-,n^+}(r_1,r_2)$, $A_{m^-,n^-}(r_1,r_2)$.

To allow inversions of DNA segments containing pairs, we consider three classes of combinations (Fig. 14):

$$B^{(1)}_{m,n}(r_1,r_2) = A_{m^+,n^+}(r_1,r_2) \cup A_{n^-,m^-}(r_1,r_2)$$

$$B^{(2)}_{m,n}(r_1,r_2) = A_{m^+,n^-}(r_1,r_2) \cup A_{n^+,m^-}(r_1,r_2)$$

$$B^{(3)}_{m,n}(r_1,r_2) = A_{m^-,n^+}(r_1,r_2) \cup A_{n^-,m^+}(r_1,r_2)$$

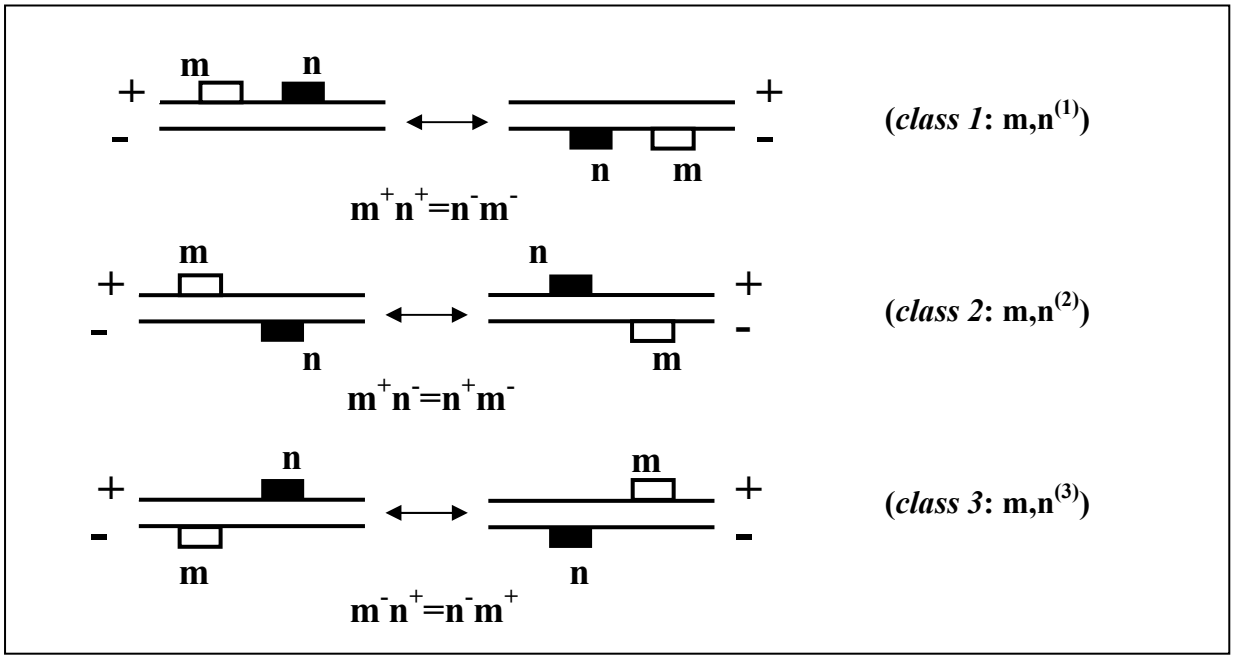


Figure 14. Pair classes. When grouping different combinations of transcription factor binding sites according to mutual orientation, we allow inversions of the whole module. This gives rise to a total of three classes as shown "m" and "n" are the binding sites for two different transcription factors; "+" and "–" denote DNA strands.

In more general form, for $i = 1,...3$ $B_{m,n}^{(i)}(r_1, r_2)$ represents the set of sequences with a pair of *i*-th class $m, n^{(i)}(r_1, r_2)$.

Let $P_t\left(B_{m,n}^{(i)}(r_1, r_2)\right)$ be a fraction of the sequences $B_{m,n}^{(i)}(r_1, r_2)$ in the (+)-training set, and $P_c\left(B_{m,n}^{(i)}(r_1, r_2)\right)$ the fraction of sequences $B_{m,n}^{(i)}(r_1, r_2)$ in the (-)-training (control) set.

We have to solve now the optimization problem to maximize the difference $P_t\left(B_{m,n}^{(i)}(r_1, r_2)\right) - P_c\left(B_{m,n}^{(i)}(r_1, r_2)\right)$ by choosing appropriate values for *m, n, i and* (r_1,r_2) (if not selected for each pair in advance using the method of distance distributions). In addition, we are interested only in pairs, which are present in at least a minimal fraction of (+)-training sequences (C_1) and in a defined maximal fraction of (-)-training sequences (C_2). They can be filtered in advance.

Thus, we search for such $B_{m,n}^{(i)}(r_1, r_2)$ for which

$$\begin{cases} P_t\left(B_{m,n}^{(i)}(r_1, r_2)\right) - P_c\left(B_{m,n}^{(i)}(r_1, r_2)\right) = \max \\ P_t\left(B_{m,n}^{(i)}(r_1, r_2)\right) \geq C_1 \\ P_c\left(B_{m,n}^{(i)}(r_1, r_2)\right) \leq C_2 \end{cases} \tag{3}$$

where $0 \leq C_{1,2} \leq 1$ are adjustable parameters.

For single pairs we chose $C_1 = 0.8$ and $C_2 = 0.4$. We could not find pairs which would satisfy more stringent parameters, i. e. either higher C_1 or lower C_2; on the other hand, requirement (3) was found to be satisfied by a lot of different combinations which gave rise to the same P_t and P_c.

To make the analysis more specific, we can consider combinations of pairs instead of single pairs. For sake of simplicity, we will omit further on (r_1, r_2) from the expression $B_{m,n}^{(i)}(r_1, r_2)$ (but it should be kept in mind that $B_{m,n}^{(i)}$ is always a function of (r_1, r_2)). Each possible type of pair is determined by values of *m, n* and *i*. We can list all types of pairs and assign a number *j* to each pair in this list. Then each type of pair is characterized by m_j, n_j, i_j:

j	*m*	*n*	*i*
1	AP-1	Elk-1	1
2	AP-1	Elk-1	2
3	AP-1	Elk-1	3
4	C/EBP	Elk-1	1
…	…	…	…

Then the sequences with the pair can be represented as $B_{m_j n_j}^{(i_j)}$. For simplicity, let us call

$$B_{m_j n_j}^{(i_j)} \equiv D_j$$

For two different j_1 and j_2 ($j_1 \neq j_2$) we can identify D_{j_1} and D_{j_2}, which appear in the (+)-training set simultaneously:

$$\begin{cases} P_t\left(D_{j_1}\right) \geq C_1 \\ P_t\left(D_{j_2}\right) \geq C_1 \\ P_t\left(D_{j_1} \cap D_{j_2}\right) \geq C_1 \\ P_c\left(D_{j_1} \cap D_{j_2}\right) \leq C_2 \\ P_t\left(D_{j_1} \cap D_{j_2}\right) - P_c\left(D_{j_1} \cap D_{j_2}\right) = \max \end{cases} \quad (4)$$

A triple or a combination of a higher order can be represented in the same way.

2.4.2. Defining complementary pairs (pairs of pairs)

The antibacterial response of the cell is triggered by at least two distinct pathways, and it may be therefore supposed that there are subsets of promoters activated by each of them. Optimally, they should be "complementary" in the sense of appearing in complementary subsets of the (+)-training set (Fig. 13).

Complementary pairs were searched first in a "seed" subset of the (+)-training set of sequences (Fig. 15, step 1). It comprises those genes for which the most reliable evidence is available that they are involved in the antibacterial response (as discussed in the subsection *Seed sequences*; see also Table 15 in *Methods* section). We considered all possible pairs, which could be found in this subset (Fig. 15, step 2). Further on, we considered all pair-wise combinations, calling pairs complementary, if:

(a) they together cover the whole subset (C_1 is therefore always set to $1, P_t\left(D_{j_1} \cup D_{j_2}\right) = 1$);

(b) each of them can be found in not more and not less than a certain number of sequences (defined by adjustable parameters C_3 and C_4, see below), with a certain allowed overlap (defined by the parameter C_5).

Thus, the requirement for complementary pairs is:

$$\begin{cases} C_3 \leq P_t\left(D_{j_1}\right) \leq C_4 \\ C_3 \leq P_t\left(D_{j_2}\right) \leq C_4 \\ P_t\left(D_{j_1} \cup D_{j_2}\right) \geq C_1 \\ P_t\left(D_{j_1} \cap D_{j_2}\right) \leq C_5 \\ P_c\left(D_{j_1} \cup D_{j_2}\right) \leq C_2 \end{cases} \quad (5)$$

where $0 \leq C_{3,4,5} \leq 1$ are adjustable parameters.

We chose $C_3 = 0.3$, $C_4 = 0.7$ and $C_5 = 0.2$. As we had no means to estimate the expected proportion of complementary pairs in the subsets, we started with these rather unrestrictive parameter settings. Finally, the chosen pairs were found in the proportion 0.4/0.6 for C_3 / C_4. In the next step, we repeated the search including the sequences that are orthologous to the "seed" set (Fig.15, step 3). We looked for those pair combinations which were found in the first step (in the human "seed" sequences). (The second and the third steps may be combined in one).

In the last step we repeated the search in the whole (+)-training set of 33 sequences, looking only for the combinations found in the second step (i.e., in the 12 "seed" and their orthologous sequences) (Fig.15, step 4).

The percentage of the pair occurrence in the (-)-training set has been counted on the first step with the subsequent filtering of pairs.

2.5. Application of the methodology

In every step of our investigations, we tried to combine purely computational approaches with the preexisting experiment-based knowledge, as it is represented in corresponding databases and literature, and with our own biological expertise. To develop a promoter model, the first task is to select those transcription factors, the binding sites of which shall constitute the model. The overwhelming majority of methods and tools estimating the relevance of predicted TF binding sites in promoter regions are based on their over- and under-representation in a positive (+) training set in comparison with some negative (-) training set. If, however, a binding site is ubiquitous, or very degenerate, so that it can be found frequently in any sequence, the comparison with basically any (-)-training would not reveal any significance of its occurrence. That tells nothing about their functionality in any specific case, which may be dependent on some additional factors and/or other conditions. Therefore, basing the decision about the relevance of a transcription factor for a certain cellular response solely on whether its predicted binding sites are over-represented in the responding promoters may lead to a loss of important information. Thus, we did not rely on this kind of evidence but rather chose the candidate transcription factors according to available experimental data.

In several steps of the model construction, we had to estimate over-representation of a feature in the (+)-training set compared with the (-)-training set. We operated with the number of sequences that possess the considered feature, in our case a pair of TFBS, at least once. Otherwise, mere enrichment of a feature in the (+)-training set may be due to strong clustering

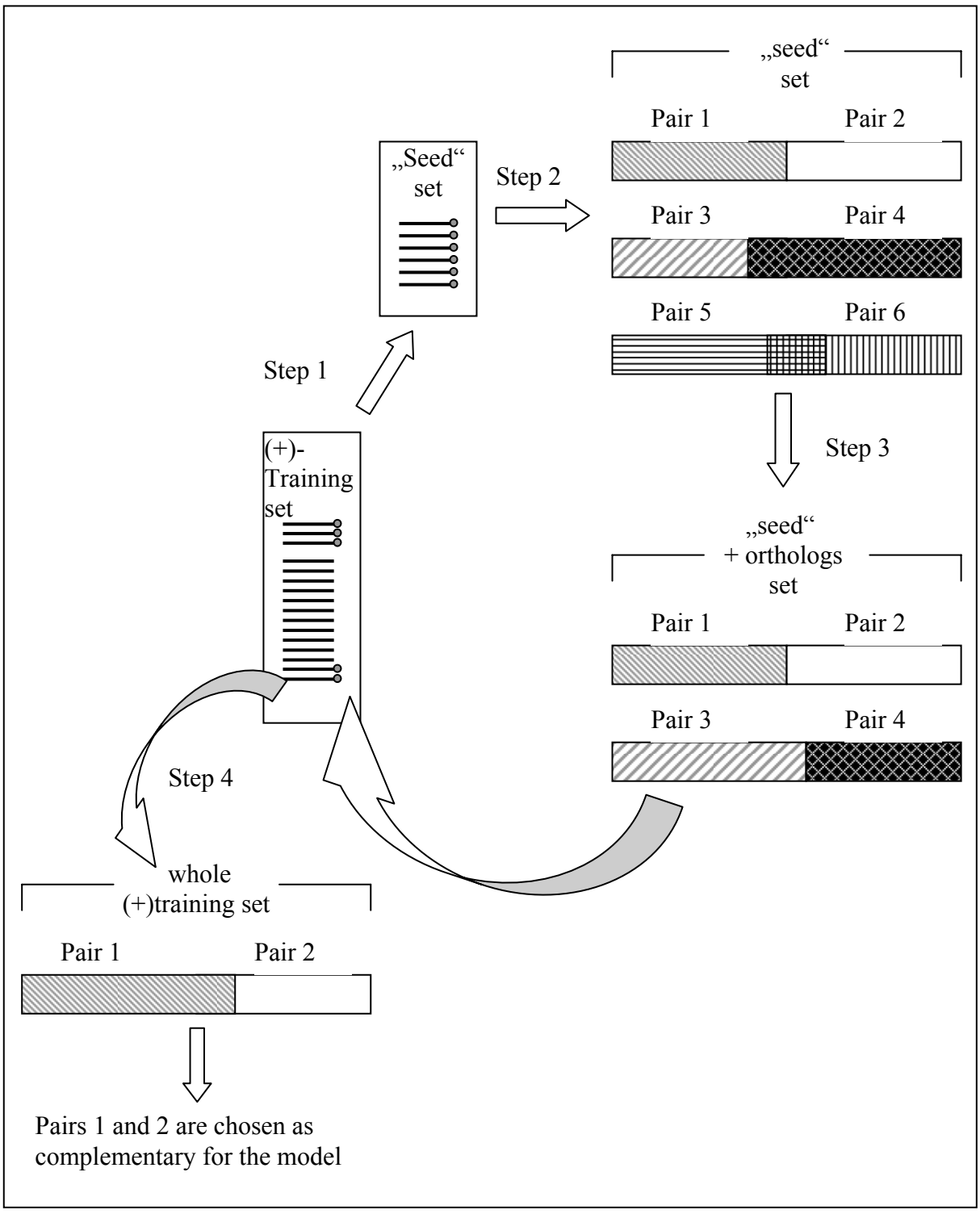


Figure 15. Algorithm of the search for complementary pairs using ”seed” sets.
Step 1. Selection of a “seed” set;
Step 2. Selection of complementary pairs in the human “seed”; every combination is checked in the (-) training set and only those, which are found in less than 40% of sequences, are taken into further consideration.
Step 3. Selection of complementary pairs in the “seed” of orthologs or in the joint “human + orthologs” “seed”. (Step 2 may be omitted and substituted by Step 3)
Step 4. Search for the selected pairs in the whole (+)-training set. After that, the final choice is made.

in a few members of that set which would not lead to a useful prediction model. At the first step the Student's t-test has been performed (the normality of distribution has been demonstrated before (data no shown)), but it appeared to be a weak filter: for example, we could find several pairs, which showed, if estimated with t-test, a remarkable over-representation ($p<0.001$), but with a difference of 97% in the (+)-training set versus 85% in the (-)-training set, which is of no practical use to construct a *predictive* model, since it is also important to have minimal occurrence of a discriminating feature in the (-)-training set. In the further work we considered all pairs with $p<0.005$, but as this did not reasonably restrict the list of considered pairs, we had to apply an additional filtering approach. For this purpose we used a simple characteristic such as the percentage of sequences in (+) - and (-)-training sets. By operating directly with percentages we could easily filter out those pairs, which would identify too many false positive sequences, thus getting rid of a substantial part of useless information. This procedure allows to estimate immediately the applicability of the model to identify further candidate genes that may be involved in the cellular response under consideration.

The main subject of this investigation is the interaction of bacterial and eukaryotic cells and the regulation of response of the latter. But the developed methods are designed to be universal, so we applied them not only to the description of the system in the main focus of this study, but also to some closely related systems. Three investigated systems involve in this or that way LPS-triggering through TLR4 receptor:

(a) epithelial cells' response to *P. aeruginosa* binding;

(b) early response of mouse monocytes to LPS triggering;

(c) MyD88-independent pathway of TLR4 triggering.

The other considered system is MALP-2 signaling pathway that is closely related to LPS signaling and is triggered through TLR2 and 6.

Promoter modeling together with the subsequent prediction of the target genes allows us to better understand one of the main pathways of the antibacterial response.

2.5.1. Epithelial cells' response to *Pseudomonas aeruginosa* binding

2.5.1.1. Selection of the "seed" set

We started our analysis with a group of "seed" sequences, which we considered for distinct reasons more reliable and preferable (the selection of the sequences for the positive training set is described in the *Methods* section; see Table 15). Choosing a seed group, we

took into consideration two kinds of evidence; the first was the source of information, i. e. the methods with which the gene has been shown to participate in the response. We took the promoter sequences of those genes which have been reported by other methods but microarray analysis (Smith *et al.,* 2001; McNamara *et al.,* 2001; Harder *et al.,* 2000; Leidal *et al.,* 2001; Kovarik *et al.,* 1996; Walsh *et al.,* 2001; Becker *et al.,* 2001), and which have been independently reported by at least two different groups.

The second kind of evidence was whether we could find any additional biological reasoning for the gene to participate in this kind of reply. For instance, a well-known participant of the NF-κB-activating pathway such as IκBα, or participants of different pathways which are likely to be triggered here as well, like c-Jun or PKC, were estimated as the first candidates for the "seed" group.

Finally, the "seed" contained 12 human sequences (Table 15). We could retrieve all mouse orthologs constituting a separate mouse "seed". We then run our analysis in either "seed" separately and in the combined human/mouse "seed", and compared the results.

2.5.1.2. Selected TFs and conditions of the search

We found 5 factors reported in the literature as taking part in anti-bacterial or similar responses and selected them as candidate TFs (Ben-Baruch *et al.,* 1995; Bergmann *et al.,* 1998; Guha and Mackman, 2001; Guha *et al.,* 2001; Gum *et al.,* 1997; Harder *et al.,* 2000; Ko *et al.,* 1997; Kovarik *et al.,* 1996; Li *et al.,* 1998; Perrais *et al.,* 2001; Smith *et al.,* 2001; Voynow *et al.,*1999; Zhang and Ghosh, 2001) (see *Methods*). Not all of these candidate TFs are over-represented in the (+)-training set used in this analysis. For instance, no over-representation has been found for important factors such as NF-κB, AP-1 and C/EBP. Nevertheless, these factors were included in the model, because not the binding sites themselves, but their combinations may be over-represented.

On the other hand, some of the factors, which have also been mentioned in the literature as potentially relevant (e.g., SRF (Heidenreich *et al.,* 1999; Dieterich *et al.,* 2003)) or which might be of a certain interest because of their participation in relevant pathways (e.g., CREB, according to the TRANSPATH database (Krull *et al.,* 2003)) were not included in the model because we could not adjust the thresholds for their detection according to our requirements (see *Methods*). For instance, SRF would be of special interest, because it is known that it tends to cooperate with Elk-1 (Dieterich *et al.,* 2003), but to identify 80% of TP we had to lower the matrix similarity threshold to 0.65, which is unacceptably low and would provide too many false positives.

Finally, we constructed our promoter model of binding sites of 5 TFs (NF-κB, C/EBP, AP-1, Elk-1, Sp1), considering their pair-wise combinations and some combinations of higher order (complementary pairs, see below). A rather large number of combinations satisfied the requirements described in the previous sections. However, when we selected those that were robust in a "leave-one-out" test for the "seed" sets, the final list of potential model constituents was shortened down to only 1 ubiquitous and 9 complementary pairs (5 characterizing one subset, 4 – the other). All materials illustrating the search process can be found in *Appendix 3a* and *-b*.

We found one satisfactory pair that should be found in all promoters of target genes:

AP-1, NF-κB$^{(1)}$(10,93)

(*AP-1, NF-κB*, class 1, distance from 10 to 93 bp; see Fig. 14 for pair classes).

The search for the combination of two or more pairs, which should be found in the whole set simultaneously, did not give any significant improvement of the results (see "Supplementary materials/Search for 2 pairs_Pa_model.doc").

Among the complementary pairs we found, several of them appeared to be interchangeable: each pair of pairs or any combination of them resulted in the selection of the same subsets from the (+)-training set (52%) (Fig. 16). Fig. 16 shows only those pairs which have been chosen for the final model, but there were several more which identified the same subset of the (+)-training set. The large number of complementary pairs may indicate that they are parts of more complex TFBS combinations, consisting of 4, 5 or more TFBS.

The false positive rate depended on the number of applied pairs; when we used all of them together, they gave only 1.7% of FP (i. e., only 1.7% of the sequences in the (-)-training set revealed the presence of all pairs under consideration). But the simultaneous usage of all the pairs could overfit the model, so we did not apply them all, sacrificing a bit of specificity for sake of a higher sensitivity.

Finally, we came up with 4 complementary pairs (Fig. 16) composed of 7 different TFBS pairs. Four of these TFBS pairs together are indicative for one subset of sequences, the remaining three for the other. As it has been mentioned before, the discovery of complementary pairs entails automatically the discovery of the corresponding subsets of sequences. We analyzed the distribution of the constituents of the found complementary pairs across the (+)-training set, which enabled us to assign the genes either to one or to the other subset, or to both (Table 9). Note that one of the subsets (subset 1) is in good agreement with the experimental data: MCP1, IL-8, β-defensin and MUC1 are known to be regulated by LPS, whereas IκBα is an important participant of this pathway; thus, these genes could be expected

to belong to one pathway and, therefore, to one subset. Here, they all belong to the subset 1. This observation provides good support for the concept of complementary pairs, which we applied here.

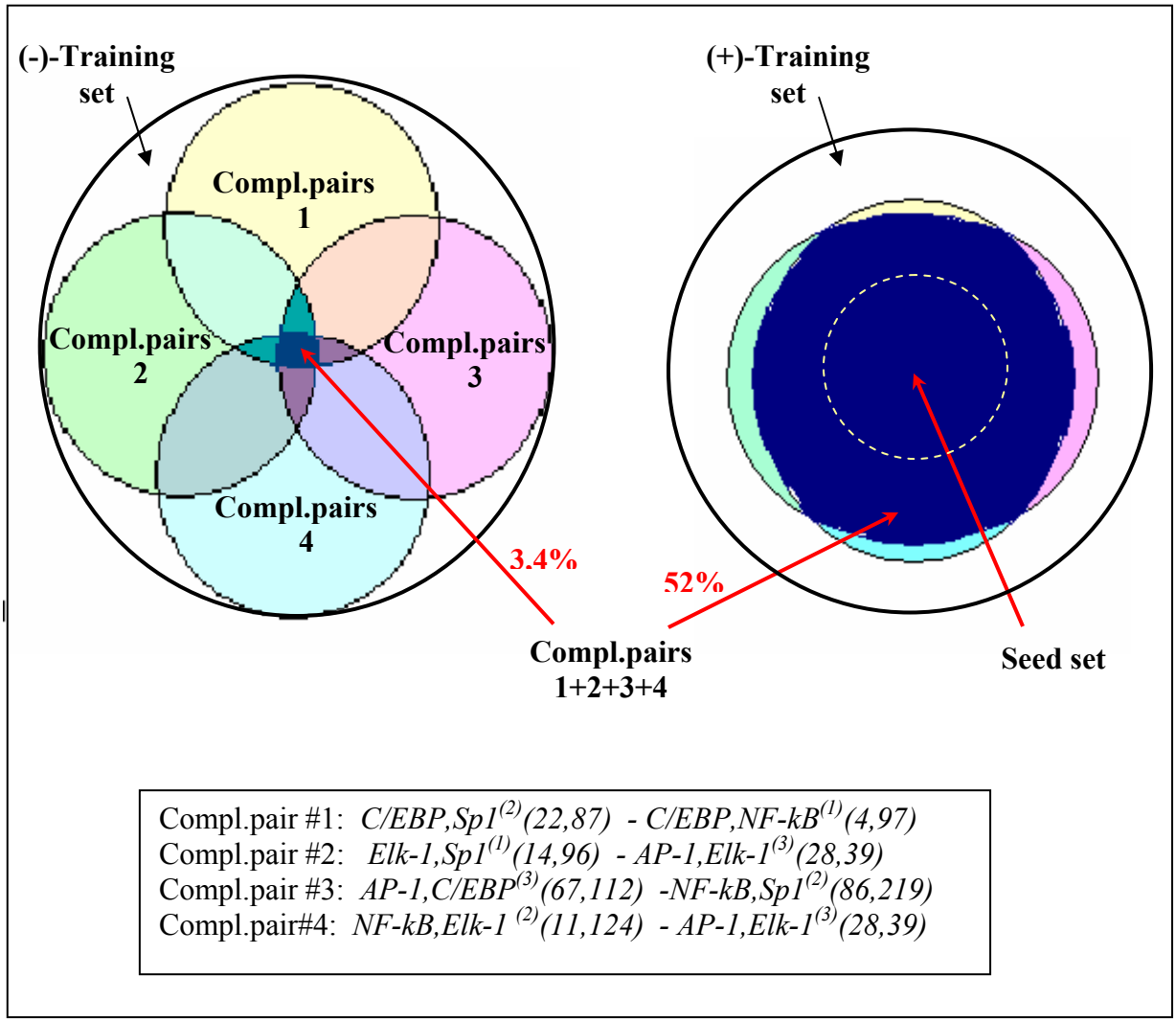


Figure 16. Seven pairs, which are combined in four complementary combinations, and the results of their simultaneous application. Each of the complementary pairs searches for nearly the same portion of the training set, while in the negative training set their intersection appears to be very small. Here, only those pairs are shown that have been chosen for the final model, but there were several more, which searched for the same subset of the training set and gave altogether 1,7% in the negative training set. Note that the circles are not exactly drawn to scale.

Table 9. Assignment of training sequences to two subsets.

	Subset 1(LPS-dependent pathway)	**Subset 2**
Complementary pairs	Elk-1,NF-κB$^{(2)}$ (11-124) Elk-1, Sp1$^{(1)}$ (14-96) C/EBP, Sp1$^{(2)}$ (22-87) AP-1,C/EBP$^{(3)}$ (67-112)	AP-1, Elk-1$^{(3)}$ (28-39) NF-κB, Sp1$^{(2)}$ (86-219) C/EBP, NF- κB$^{(1)}$ (4-97)
Regulated genes (in the training set)	MCP1* IL8* β-Defensin* MUC1* ELF3 cytochrome p450 IkBα*	PKC, proteinkinase C TEL2 c-jun (?) TFPI-2
	RhoB, PLAU, IRF-1, hORC2L	
Not assigned	SLC29, DPH2L2, FPGS,	

Genes marked with asterisk are known to be activated through LPS-dependent pathway; note that they all belong to one subset.

In order to avoid the over-fitting of the model and to demonstrate the significance of our results, we performed a permutation test. For that, we conducted 2000 iterations of random permutation of (+) and (-) labels in the training sets and tried to rebuild the model using the procedure described above. The rate of correct classification on this random selection was estimated. The cases of common and complementary pairs were considered separately. The analysis was made for different C_1, C_2 ($0.7<C_1<0.8$, $0.4<C_2<0.5$) for common pairs; for complementary pairs we considered the case with C_3=0.3, C4=0.7 and C_5=0.2. The probability to find by chance a "seed" of 12 sequences which would produce at least one pair common for the random selection of 33 sequences (including the "seed") depends on the chosen C_1, C_2 and is found to vary between $p<0.0005$ (C_1=0.8, $C2$=0.4, the parameters used for our model construction) and p=0.02 (C_1=0.7, $C2$=0.4). We failed to find any complementary pairs after 1000 iterations of the permutation test with the parameters used for the "real" (not permuted) model construction. These results suggest that the success of the model construction based on the search for combinations of TFBS is strictly dependent on the selected training set (thus, on our prior biological knowledge) and that the significance of the findings, depending on the correct choice of the adjustable parameters, is high enough to claim their non-randomness. Thus, we can say that in the described case the pairs found in the given (+)-training set with the given parameters are the real characteristics of this set.

2.5.1.3. Promoter model

The model called "P.a.-model" (for *Pseudomonas aeruginosa*) consists of two kinds of combinations of pairs: ubiquitous pairs (which should be found in all promoters of the target genes), and complementary pairs. We can divide the model into two modules, one for each kind of combination.

Let *M*1 and *M*2 be modules comprising ubiquitous pairs and complementary pairs, respectively.

Module *M1* comprises the pair $AP-1, NF-\kappa B^{(1)}(10,93)$.

Module *M2* comprises all complementary combinations listed in the Fig. 16. Each complementary pair can be taken as a submodule (*m*) in *M2*.

To apply the model means to search for sequences containing all these combinations. Let us call *S(M)* the set of sequences which possess the whole model *M*; then we can also consider *S(M1)* and *S(M2)* (the sets possessing the modules *M1* and *M2*, respectively), and *S(m)* – the set with a submodule *m*.

Then

$$S(M1)=B^{(1)}_{AP-1,NF-kB}(10,93)$$

Module M2 consists of submodules (*m*); in this case we consider four submodules, so the sequences containing M2 can be found as:

$$S(M2)=S(m_1)\cap S(m_2)\cap S(m_3)\cap S(m_4)\ ,$$

where the set with each submodule we must consider as a union of sequence sets containing the complementary pairs:

$$S(m_1)=B^{(2)}_{C/EBP,Sp1}(22,87)\cup B^{(1)}_{C/EPB,NF-kB}(4,97)$$

$$S(m_2)=B^{(1)}_{Elk-1,Sp1}(14,96)\cup B^{(3)}_{AP-1,Elk-1}(28,39)$$

$$S(m_2)=B^{(3)}_{AP-1,C/EBP}(67,112)\cup B^{(1)}_{NF-kB,Sp1}(86,219)$$

$$S(m_3)=B^{(2)}_{Elk-1,NF-\kappa B}(11,124)\cup B^{(3)}_{Ap-1,Elk-1}(28,39)$$

The final result of application of the model M can be presented as

$$S(M)=S(M1)\cap S(M2)$$

The model gives 3.4% of false positives and re-identifies 52% of the whole (+)-training set, but these 52% comprise all most reliable sequences of the set (remember that we must allow for some reduction because the set is not absolutely reliable).

2.5.1.4. Identification of potential target genes

Applying our promoter model to screening of 13000 upstream regions from a collection of human 5'-flanking sequences (Kel-Margoulis *et al.,* 2003), we identified about 580 genes as harbouring this combination of TFBS. After erasing all those that encode hypothetical products, we came up with a list of 430 potential target genes, which can be checked for plausibility. More than 60% of these genes encode different representatives of the immune system, which can be expected to participate in the cells response, as well as transcription factors and other regulatory proteins. Some of the most interesting potential target genes are shown on the Table 10. The whole data set can be found in the "Supplementary materials".

Table 10. Selection of candidate genes identified by the promoter model.

TNFRSF14 tumor necrosis factor receptor superfamily TNFAIP6 tumor necrosis fact., alpha-induced protein 6 PPP3CA protein phosphatase 3 (calcineurin A) NLI-IF nuclear LIM interactor-interacting factor WISP1 WNT1 inducible signaling pathway protein 1 IL8 interleukin 8 TFPI2 tissue factor pathway inhibitor 2 DEFB2 defensin, β2 POU2F1 POU domain, class 2, transcription factor 1 MAP2K1IP1 MAPKK1 interacting protein 1 CSF2 colony stimulating factor 2 (granulocyte-macrophage) TAF2F TATA box binding protein (TBP)-associated factor RNA polymerase II, F, 55kD ABT1 TATA-binding protein-binding protein CALN1 calneuron 1 TRAF1 TNF receptor-associated factor 1 FPGS folylpolyglutamate synthase RENT2 regulator of nonsense transcripts 2 CYP26A1 cytochrome P450, subfamily XXVIA EHF ets homologous factor, MAP3K11 mitogen-activated prot. kinase kin. kinase 11 IRAK-M interleukin-1 receptor-associated kinase M ARHGDIA Rho GDP dissociation inhibitor (GDI) α HSY11339 GalNAc alpha-2, 6-sialyltransferase I, long form	HCNGP transcriptional regulator protein CYP4F11 cytochrome P450, subfamily IVF IRF3 interferonregulatory factor 3 ICAM3 intercellular adhesion molecule 3 PPARA peroxisome proliferative activated receptor, α IKBKG inhibitor of κ light polypeptide gene enhancer in B-cells, kinase γ ELK1 ELK1, member of ETS oncogene family STK31 serine/threonine kinase 31 SERPING1 serine (or cysteine) proteinase inhibitor GPR4 G protein-coupled receptor 4 RAB5B RAB5B, member RAS oncogene family RAB7 RAB7, member RAS oncogene family NFKB1 nuclear factor of κ light polypeptide gene enhancer in B-cells NFKBIB nuclear factor of κ light polypeptide gene enhancer in B-cells inhibitor, beta CEBPE CCAAT/enhancer binding protein (C/EBP),ε ELK1 ELK1, member of ETS oncogene family EHF ets homologous factor *15 Zinc finger proteins* *small inducible cytokine subfamily A (Cys- Cys), members 5,11, 20 and 23* *Interleukins:* IL1, IL1delta, IL8, IL12A, IL12B, IL13, IL23

The whole list can be found in "Supplementary materials/Pa_model_potential target genes 1", "-2".

2.5.2. LPS triggering: promoter model for immediate early response

2.5.2.1. Selection of the relevant TFs

An extensive literature search gave us the selection of 10 TFs as relevant for the triggering of the LPS-responsive genes (see *Methods*). The set of the transcription factors included: AP-1, ETS, Elk-1, NF-κB, ATF2, C/EBP, CREB, NFAT, Sp1 and SRF. In this part

of work we decided to reduce the lowest possible threshold for PWM and accepted the thresholds which gave 80% of the TP rate but which were not lower than 0.80/0.65 (core similarity/ matrix similarity). This allowed us to include in the search CREB and SRF that are well known to play important role in RSK pathway.

The positive training set of sequences was based on the data published by C.Scheidereit and coworkers (Krappmann *et al.,* 2004) (see *Methods*). Only the genes that performed more than 2,5-fold increase of induction after 90 min of LPS treatment were chosen for the set.

2.5.2.2. Search for combinations

We started the search for single pairs with the "distance distribution" approach, filtering out all pairs which did not demonstrate any over-represented distances. The result of the search is shown on Table 11.

Table 11. Pairs found on over-represented distances in the set "90_LPS", which comprises 90 sequences of genes triggered within 90 minutes after the LPS treatment (see *Methods*).

	AP-1	**Sp1**	**CREB**	**NFAT**	**ATF-2**	**SRF**	**ETS**	**Elk-1**	**C/EBP**	**NF-κB**
AP-1		+		+						+
Sp1			+			+				+
CREB				+		+				+
NFAT						+				
ATF-2										
SRF										+
ETS										
Elk-1										+
C/EBP										

Mutual orientation is not considered.

Every pair from the Table 11 was checked for the number of sequences in which it occurred (Table 12). Only the pairs present in more than 65% of the TP set were taken into further consideration (the selection of the percentage will be considered in the *Discussion*).

Table 12. Pairs covering more than 65% of the TP set.

Pair	**Distances**	**% in TP**
AP-1 - Sp1	**26-78**	**90**
AP-1 - NFAT	**8-34**	**87**
AP-1 - NF-κB	**8-24** **81-112**	**63** **82 (8-112)**
Sp1 - CREB	**31-76**	**73**
CREB - NFAT	**9-23**	**67**

The pairs found in more than 65% of the TP set were combined with each other first pair-wise, selecting the best combination (i.e., present in the highest number of sequences). Then to the selected pair of pairs we added the next pairs (also selecting the best combination).

Finally, we chose three pairs (AP-1 – Sp1, AP-1 – NFAT, AP-1 – NF-κB) which could be combined in the optimal way (Fig. 17). Two other two pairs (CREB – Sp1 and CREB – NFAT) were considered as next candidates (see *Appendix 4*).

Different combinations of pairs are demonstrated on Fig.17A. Pair-wise combinations of (AP-1 – Sp1 and AP-1 – NF-κB), (AP-1 – Sp1 and AP-1 – NFAT) and (AP-1 – NFAT and AP-1 – NF-κB) cover 77%, 77% and 73% of the TP set, respectively. When combined together, the two best pairs of pairs ((AP-1 – Sp1 and AP-1 – NF-κB), (AP-1 – Sp1 and AP-1 – NFAT) give 68% of the TP set. As all these pairs contain AP-1, it is reasonable to suppose that they can belong to a triple (or higher order) combination sharing the AP-1 binding site. Thus, we checked different possible triple combinations which could be constructed from these pairs. The best two triples covered each 74% of the TP set (Fig. 17B). When considered simultaneously, these triples covered only 57% of the TP set. Nevertheless, it is still more than a half of the set, and deserves consideration. Making the next step forward, we asked ourselves, whether the pair-wise combinations sharing the AP-1 site (shown in Fig. 17A) could be organized in a quadruple. Having no special programs for quadruples, we added other triples which could be expected in a quadruple constructed from the binding sites for AP-1, Sp1, NFAT and NF-κB. The best combination appeared to be NFAT – Sp1 - NF-κB. When combined with the other two triples, it covered 53% of the TP set. Thus, we can suppose that there really exists the quadruple combination, although it can be found in only slightly more than a half of all sequences.

The other possibility we considered as an alternative to the quadruple combination was the combination of the found triples and some other pairs (those called above "the next candidates": CREB – Sp1 and CREB – NFAT). These pairs demonstrated a good percentage of the TP set, as well (see Table 12), but their combination with the complex of triples dramatically dropped the percentage of re-identification. We decided to look at the combinations of these pairs with single triples. The best result was achieved for the combination of the triple AP-1 – Sp1 and AP-1 – NF-κB and the pair CREB – Sp1 (Fig. 17D)

We could not find any complementary pairs.

2.5.2.3. Promoter models

We constructed several descriptive models of the regulatory pattern(s) which are probably used in the early LPS-triggered response. No one combination can be used as a predictive model, since they all give too high rate of the false positives.

Here we show three promoter models which give approximately the same result (51, 52 or

53% of the TP rate, 9, 13 or 12% of the FP rate). Other promoter models can be constructed from the combinations shown on Fig 17.

Model A:

Contains three "common" triples (numbers in parentheses show the distance; the location on DNA strands (+ or -) is not considered here):

- NF-κB (8-112) Sp1 (26-130) AP-1 (the order of the elements can be as shown or reverse: AP-1 – Sp1 - NF-κB with the same ranges of distances; the both variants are searched with an "or" operator);
- AP-1 (8-34) NFAT (8-96) Sp1 or NFAT (8-34) AP-1 (26-78) Sp1 (the both variants are searched with an "or" operator);
- NF-κB (8-112) Sp1 (26-120) NFAT or the reverse (NFAT- Sp1 - NF-κB with the same ranges of distances).

Let us denotate a subset of sequences containing a triple combination of the binding sites *m, n* and *l* as

$T_{m(r_1,r_2)n(r_3,r_4)l}$,

where r_1,r_2 and r_3,r_4 are the distances between *m* and *n* and *n* and *l*, respectively.

Using the denotations from 2.5.1.2., we can describe the promoter model as follows:

$$S(m_1) = T_{NF-\kappa B(8-112)Sp1(26-130)AP-1} \cup T_{AP-1(8-112)Sp1(26-130)NF-\kappa B}$$

$$S(m_2) = T_{AP-1(8-34)NFAT(8-96)Sp1} \cup T_{NFAT(8-34)AP-1(26-78)Sp1}$$

$$S(m_3) = T_{NF-\kappa B(8-112)Sp1(26-120)NFAT} \cup T_{NFAT(26-120)Sp1(8-112)NF-\kappa B}$$

$$S(M) = S(m_1) \cap S(m_2) \cap S(m_3)$$

This model gives 53% of TP and 13% of the FP (of the control set) (Fig. 17C)

Model B:

Contains two "common" triples and a "common" pair:

$$S(m_1) = T_{NF-\kappa B(8-112)Sp1(26-130)AP-1} \cup T_{AP-1(8-112)Sp1(26-130)NF-\kappa B}$$

$$S(m_2) = T_{AP-1(8-34)NFAT(8-96)Sp1} \cup T_{NFAT(8-34)AP-1(26-78)Sp1}$$

$$S(m_3) = B^{(1)}_{Sp1,CREB}(31,76) \cup B^{(3)}_{CREB,Sp1}(31,76).$$

$$S(M) = S(m_1) \cap S(m_2) \cap S(m_3)$$

This model re-identifies 52% of TP and gives 12% of the FP rate.

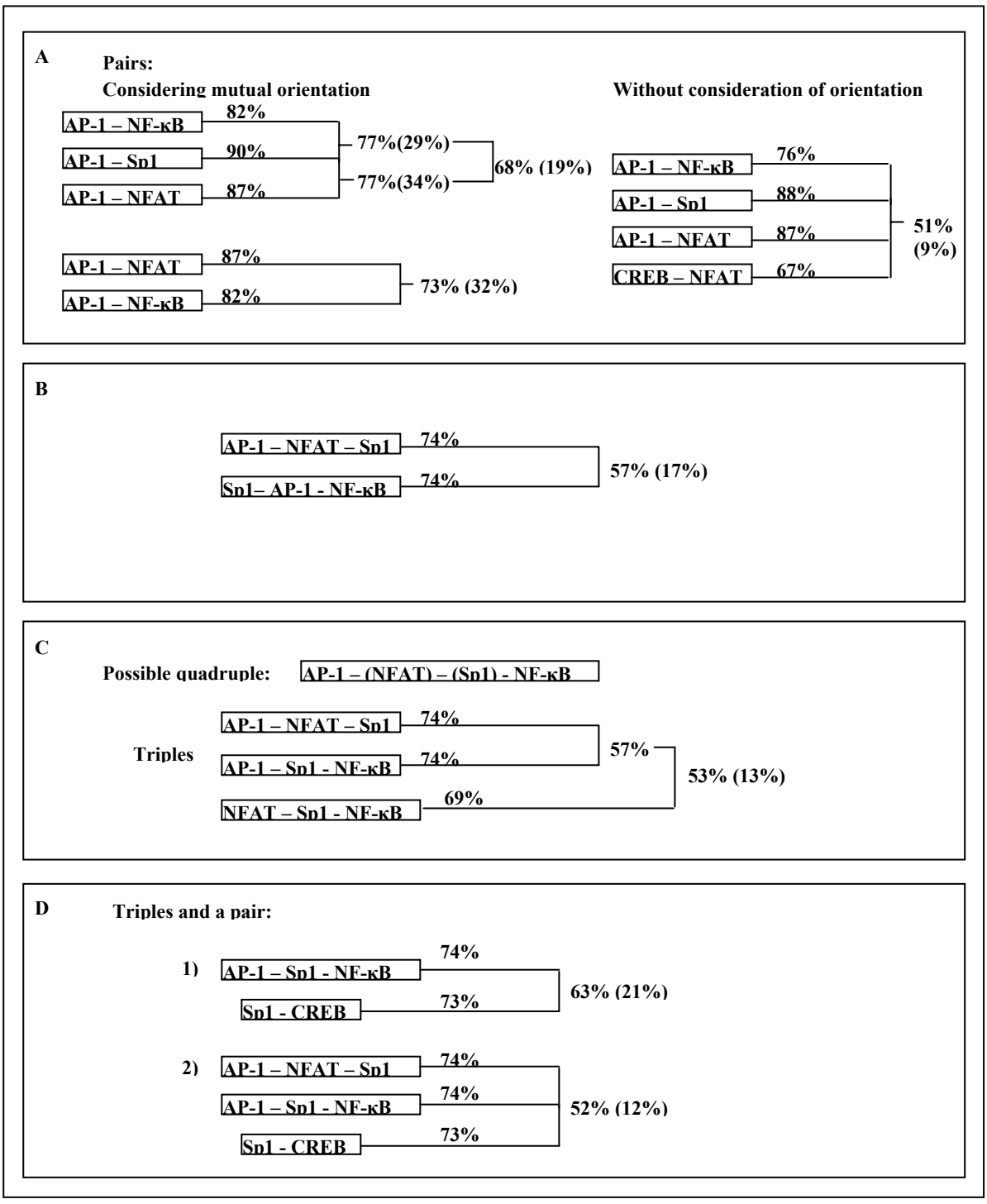


Figure 17. Combinations of pairs. Numbers show the percentage of sequences containing a pair or a triple in the TP set; numbers in parentheses show the percentages in the control set.

A. Pair-wise combinations of (AP-1 – Sp1 and AP-1 – NF-κB), (AP-1 – Sp1 and AP-1 – NFAT) and (AP-1 – NFAT and AP-1 – NF-κB) cover 77%, 77% and 73% of the TP set, respectively. When combined together, (AP-1 – Sp1 and AP-1 – NF-κB) and (AP-1 – Sp1 and AP-1 – NFAT) give 68% of the TP set.
B. Triple combinations which can be constructed from the pairs from (A). The best two triples cover each 74% of the TP set. When considered simultaneously, these triples covered only 57% of the TP set.
C. Possible quadruple combination. Having no special programs for quadruples, we checked other triples which could be expected in a quadruple constructed from the binding sites for AP-1, Sp1, NFAT and NF-κB. The best additional triple appeared to be NFAT – Sp1 - NF-κB. When combined with the other two triples, it covered 53% of the TP set.
D. An alternative (to (C)) combination of a triple and a pair. 1) The best variant covers 63% of the TP set. 2) Two triples and a pair give 52% of the TP, but a better result for the control set (12%).

Model C:

The third possible combination of TFBS which can be suggested as a promoter model consists of the binding sites for NF-κB, AP-1, NFAT, Sp1 and CREB taken without consideration of mutual orientation; the four pairs should be taken with "and" operator:

$$S(M) = B_{AP-1,NFAT}(8,34) \cap B_{AP-1,Sp1}(26,78) \cap B_{CREB,NFAT}(9-23) \cap B_{NF-\kappa B,AP-1}(8-112)$$

This combination gives 51% of TP rate and 8,6% of the false positives.

As it has been mentioned above, we cannot suggest any of these models as a predictive model because of relatively high rates of the false positive predictions. To polish the models, to select the best and to avoid the risk of over-fitting we need more initial experimental data. We will return to this in *Discussion*.

2.5.3. MyD88-dependent and -independent pathways in TLR4 triggering

2.5.3.1. Promoter model for MyD88–independent pathway. Re-identification of the NF-kappaB/IRF composite element as playing the main role in the regulation of this pathway.

TLR4 signaling may occur through pathways that are either dependent on or independent of MyD88, a general adaptor protein for interleukin-1- and toll-like receptors (Kawai *et al.,* 1999). The main idea of the application described in the following was to check if we can separate the two subsets (for MyD88-dependent and –independent pathways) using the approach of the complementary pairs.

No data are published about a large-scale experiment, which would directly compare the two pathways. Nevertheless, there are sufficient data about the IRF3-dependent genes triggered through TLR4 as well as about the genes triggered by the "standard" pathway through MyD88 (see *Introduction*, 1.2.3.4., and Fig.2). The problem is that those genes, which are triggered through the MyD88-dependent pathway, may also be a subject to the MyD88-independent triggering; the fact that they belong to one pathway does not exclude that they can belong simultaneously to the other. To make our experiment more pure, we decided to use a joint set consisting of: (i) the genes, which are subject to MyD88-independent triggering (i.e., IRF3-responsive) and (ii) the genes that are not triggered by this pathway. For this purpose we took MALP-2-responsive genes. The MALP-2 signaling pathway does not have a MyD88-independent subpathway and does not include any IRF triggering. Thus, MALP-2 pathway can be used as a model of "ideal" MyD88-dependent pathway. This "substitution" of TLR4-triggered MyD88-dependent pathway with the MALP-2-triggered pathway is possible because these pathways are practically identical in the MyD88-dependent

part (Kawai *et al.,* 2001; Sato *et al.,* 2000).

Thus, we constructed an artificial training set consisting of:

a. Set of MALP-2-responsive genes
b. IRF-responsive genes (taken from literature and TRANSFAC database, see *Methods*)

Applying the method of complementary pairs along with the approach of distance distributions, we could define three patterns that successfully re-identified the input subsets.

2.5.3.2. Re-identification of the MALP-2 subset

The search with complementary pairs approach gave 6 kinds of pairs characterizing the MALP subset (see *Appendix 5a*):

Elk-1 – SRF, C/EBP – CREB, CREB – AP-1, CREB – ETS, CREB – SRF.

The distance distribution approach applied to the MALP-subset resulted in several over-represented pairs, which are presented in the Table 13A.

Only 1 pair was found with the both approaches (CREB – SRF). Nevertheless, as we used rather strong constraints for the both approaches, there was a possibility that we overlooked some useful pairs. Thus, we decided to check the other pairs found with the distance distribution approach.

Table 13. Over-represented distances in MALP subset.

A

	AP-1	ETS	SRF	Elk-1	CREB	Sp1	ATF-2	NF-κB	NFAT	C/EBP
AP-1			+					+	+	
ETS								+		
SRF					++			+	+	
Elk-1								+		
CREB						+				
Sp1									+	
ATF-2										
NF-κB										
NFAT										+
C/EBP										

B

	% of TP
SRF-CREB	**60**
SRF- NF-κB	**70**
Elk-1- NF-κB	**67**
SRF-NFAT	**77**
AP-1-NFAT	**67**

A. All over-represented pairs found in the MALP subset. Double "plus": the pair found also with the complementary pairs approach.
B. Pairs covering more than 50% of the MALP subset.

First of all, we considered only those pairs which were found in more than 50% of the MALP subset (Table 13B). The reason for the selection of the 50% is the same as in the case with LPS-triggered genes (see 2.5.2.2.): we cannot presuppose that all genes are absolutely reliable and that those that are belong to one pathway. Moreover, they can be a subject to differential regulation, too (see Fig. 4 in *Introduction*). Thus, here we also consider the pairs occurring in >50% of sequences.

From the 5 pairs conforming this demand only 3 (SRF- NF-κB and SRF – NFAT and SRF - CREB) could be combined together so that the whole combination was found still in 60% of the subset.

The pattern for the MALP-2-responsive genes consists of three combinations:

a. SRF-NF-κB (5-54), orientation 1 and 3;
b. SRF-NFAT (5-34; 55-80) (two distance peaks); orientation 1 and 3;
c. SRF-CREB (28-38; 60-72) (two distance peaks); orientation 2 and 3.

This combination re-identifies 60% of TP and 13% of FP. Thus, it cannot be used as predictive model (for predictions of potential target genes): the rate of false positives is too high. Nevertheless, it gives a clear picture of what is important in this regulatory module, and can be used as a descriptive model.

The plots illustrating the over-represented distances for the selected pairs are shown in Fig. 18. The other supportive material, which allows to see how the selection has been made, one can find in *Appendix 5b*.

Note that three transcription factors (NF-κB, CREB and SRF) involved in this pattern belong to one subpathway, namely going through P90RSK (see scheme in Fig. 4 in *Introduction*). This pathway seems to be especially important because it involves the phosphorylation of CREB, which is connected with the simultaneous phosphorylation of histone H3. We will return to this in more detail in the *Discussion*.

2.5.3.3. Re-identification of the IRF subset

The subset of sequences of the IRF-responsive genes was clearly separated from the MALP-dependent sequences by having a combination of IRF-3 - NF-κB binding sites (see tables in Appendix *5a*). The robustness of the composite element IRF- NF-κB was checked with the "leave-1-out" test (see *Appendix 5c*).

The other pairs found as characterizing the IRF subset were:

AP–1 - IRF, ETS – IRF, ATF-2 – IRF, CREB – IRF, NFAT – IRF, Elk-1 – IRF.

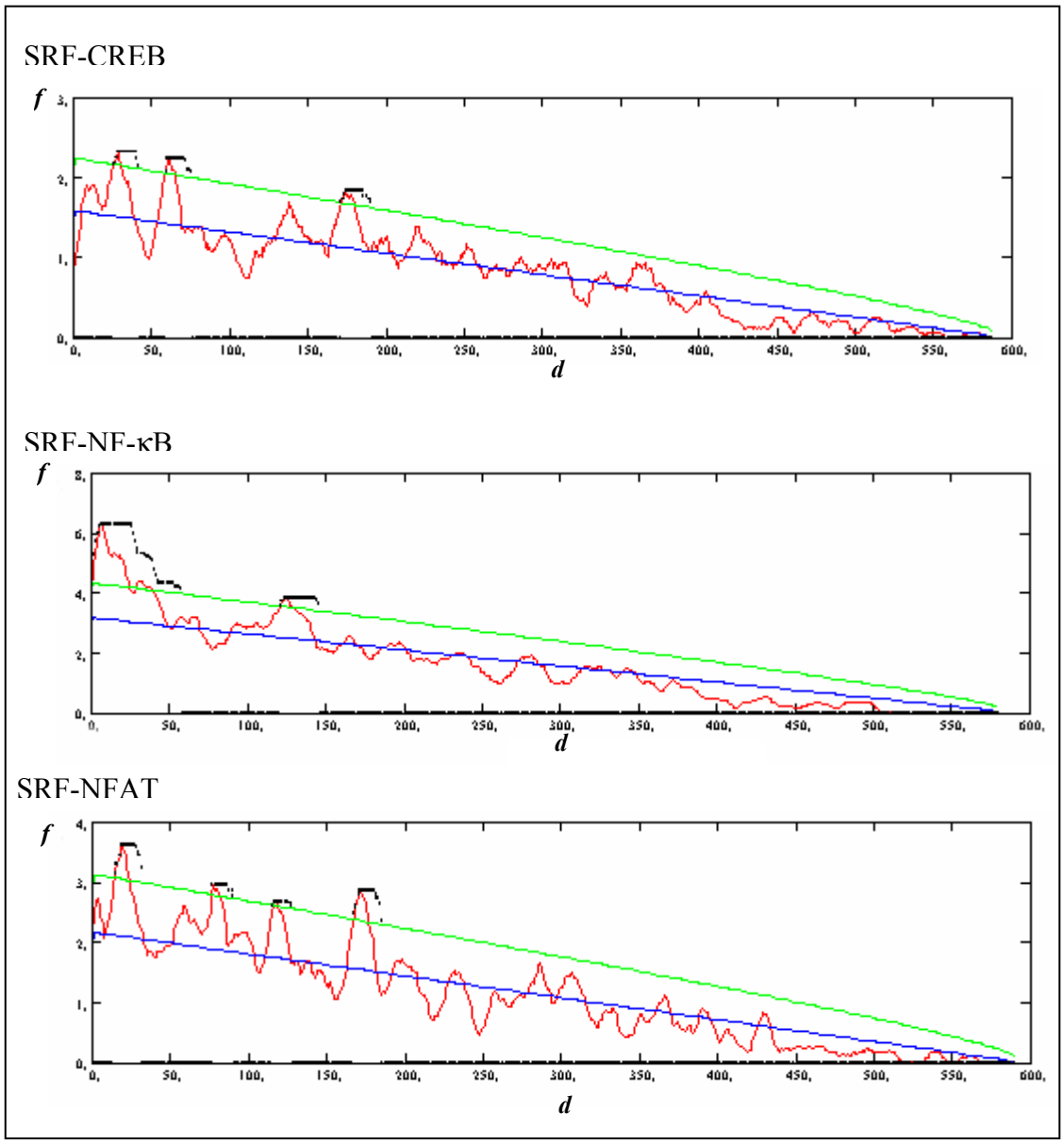


Figure 18. Three pairs selected for the promoter model for MALP-2-responsive genes demonstrate clear over-representation of some characteristic distances. Red line - results of measurement of the average number of pairs per sequence f ($f_{d,\delta}$) normalized by the number of sequences at distance d; blue line – calculated random distance distribution plus 3 standard deviations (green line). For the over-represented distances, the peaks are shown with δ (black line above the peaks) (see explanations for Fig. 11)

Using the distance distribution approach, we could identify several pairs which were of potential interest as characteristics of the IRF-responsive subset (Table 14, A). Note that three pairs were identified with the both approaches (AP–1 - IRF, CREB – IRF, NFAT – IRF). Two of these potentially most interesting and reliable pairs were found in more than 80% of the MALP subset.

Then we checked which of the found pairs were present in more than 80% of sequences of the IRF subset (Table 14, B). We were interested in the pairs, which cover the maximal number of sequences; the initial hypothesis was that all these sequences belong to one pathway, and we wanted to confirm this idea. Thus, we were not interested in sub-dividing the subset into smaller parts, which would be also statistically unreliable (the IRF subset is very small: only 11 sequences), and considered the pairs found in >80% of the set.

Table 14. Over-represented distances in IRF-subset.

A

	AP-1	**SRF**	**ETS**	**NFAT**	**CREB**	**C/EBP**	**ATF-2**	**Elk-1**	**IRF**	**NF-κB**
AP-1									++	+
SRF				+		+	+		+	
ETS							+			
NFAT									++	+
CREB										+
C/EBP										
ATF-2										+
Elk-1										+
IRF										++
NF-κB										

B

	% of TP
IRF- NF-κB	**91**
NFAT - IRF	**100**

A. All pairs found on the over-represented distances in the IRF-subset. Double "plus": pairs found also with the complementary pairs approach.
B. Pairs covering more than 80% of the IRF-subset.

The both pairs conforming this demand (IRF- NF-κB and NFAT – IRF, Fig. 19) could be combined together without loosing the representativity in the training set. These two pairs were selected for the promoter model:

IRF- NF-κB (25 – 87) orientation 1 and 2;

NFAT – IRF (5 - 44), orientation 1 and 2.

The combination of these pairs re-identifies 100% of the "IRF part" of the training set and 8 from 9 known IRF3 - NF-κB composite elements (89%). It gives 2.1% of false positive rate.

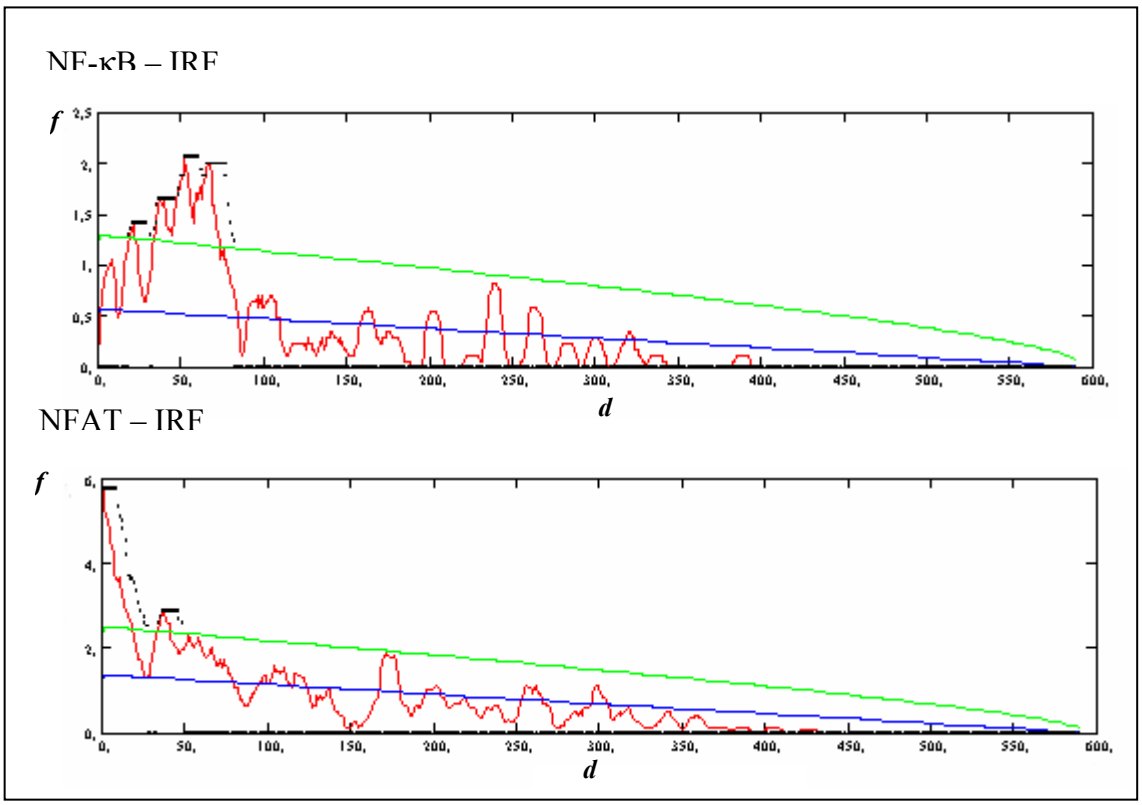


Figure 19. Two pairs characterizing the IRF-responsive subset.

Note that the combination NF-κB – IRF-3 (25-86) looks exactly like the known composite elements of NF-κB - IRF type. From 10 sequences used for this study as the "IRF-responsive part" only two (IL-15 and IFN-β) were known to have the proven NF-κB-IRF composite element. We found the described pattern in 9 from the 10 sequences. Only 2 sequences from 48 in the MALP-dependent subset were found to contain the NF-κB-IRF composite element (see *Discussion*).

3. DISCUSSION

Understanding and modeling of genetic networks is impossible without knowledge about the key participants of the regulatory process: transcription factors and regulated genes, more particularly their regulatory regions. We can consider the gene regulatory process on different levels, but the essence of it is what happens in the immediate surrounding of the gene, both in temporal and in spatial sense.

The process of transcription regulation is conducted by the interaction of TFs with specific binding sites on regulatory sequences (TFBS). The number of known (experimentally proven) TFBS can be estimated as <1% of the total number of expected TFBSs. With such a poor coverage of what we expect in reality, it is impossible to construct a comprehensive gene regulatory network (GRN). Thus, we have to include not only proven, but also predicted nodes in the GRN, i.e., predicted TFBS. The reliability of predictions of potential single TFBS is still very low; therefore, we have to add some other characteristics, such as distances between the TFBS or their orientation, thus describing a promoter as a whole. We call such a specific combination of sequence elements modulating transcription and characterizing a promoter a promoter model.

This work describes the development of methods for promoter model construction. The structure of specific promoters possesses vast information about the ways how transcriptional regulation is achieved and about ascending as well as descending pathways of the corresponding signaling. Identification of characteristic TFBS shows the role of specific transcription factors in triggering the specific response, which in turn sheds light on the signaling pathways activating these transcription factors. Detection of specific location of the TFBS, including their relative distances and mutual orientation, enables to construct promoter models which describe the characteristic promoter features of a certain gene or a group of co-regulated genes. Application of such specific models to the search in the promoters of other genes allows to predict which of these genes can be potentially involved in the same cellular response. In many cases, the products of these genes are participants of the signaling pathway that has triggered their own expression or may influence the steps of the next signaling pathways. Our concept of the circuit of interactions, therefore, enlarges and becomes more detailed.

The newly developed methods were applied to the construction of promoter models for four defensive systems of eukaryotic cells. The obtained models enabled us to get a better insight into the pathways of the corresponding signaling networks.

3.1. Development of methods

3.1.1. Subtractive approach to matrix generation

Transcriptional regulation occurs mainly through the binding of transcription factors to their binding sites in the regulatory regions. Thus, the TF binding sites can be considered as elementary units of regulatory patterns, and the detection of regulatory patterns depends on the detection of these elementary units. The quality of the recognition or prediction of TFBS is still crucial for the prediction of the whole regulatory module. The first part of this work is devoted to the improvement of the quality of the search for potential binding sites with PWMs through application of a set of matrices instead of one general matrix for a given TF. The basic idea is that sometimes the whole pool of reported experimentally proven binding sites for one transcription factor may represent a mixture of different subclasses of binding sites, serving as docking sites for different subgroups of binding factors (for instance, different isoforms of one factor, or heterodimers with distinct co-factors). In such a case, any attempt to derive a PWM from all available sequences of binding sites will lead to an extremely weak matrix. On the other hand, every binding sub-pattern can be described by its own PWM, and to detect potential binding sites one has to apply these PWMs simultaneously. The advantage is that in such case every PWM of the set can be applied with a significantly higher threshold than a single matrix applied alone. As a consequence, the number of false positive predictions is reduced.

The suggested novel approach is called "subtractive" because of the step-wise reduction of number of sequences in the training set which are used for the motif extraction and the subsequent matrix generation. Thus, the first matrix always represents the majority of the binding sites of the training set; the last will point at some minor group. There is a inherent danger in these minor groups, because it may well be that they actually represent a binding site of some other transcription factor in the vicinity of the site of interest (which is especially likely when this site is a constituent of a composite element) or some other specific sequence element, which even may have nothing to do with transcription regulation. To avoid such situations, we restricted the length of the considered sequences and required a certain minimal size of a sequence set to be used for matrix construction. This number should be determined specifically, depending on the binding site, the researcher and his specific demands. In the case of our application (C/EBP matrices) we decided not to consider groups with less than 10 sequences. But we assume that there may be situations when a matrix made of an even smaller number of sites can gain some interest. The length of the sequences we considered in

our application was the length of the binding site as it is reported in the TRANSFAC® database plus 8 flanking base pairs. This restriction is not enough to warrant it free from additional binding sites, especially if there are overlapping ones. But the prolongation of the sequences is necessary to allow small shifts of the coordinates, because we cannot be absolutely sure in the exact location of the reported binding sites; a range of 8 bp was chosen because this corresponds to the average length of a standard consensus sequence of this (and most other) binding site(s).

The application of this approach to the improvement of the situation with C/EBP PWMs demonstrated its usefulness. As a result we have obtained the set of matrices which identify complementary subsets of sequences of the training set and, being applied simultaneously, allow to reduce the rate of false positives by more than two thirds compared with the standard matrices of TRANSFAC.

We classified the sequences of C/EBP binding sites in two ways: (i) by a sequence-analysis based subtractive approach and (ii) through a functional classification, the only functional group has been identified being the C/EBP sites of the composite elements with NF-κB. It turned out that the first pattern derived from the subtractive approach was nearly the same as that constructed from the functional class of composite elements, thus confirming that the C/EBP sites within composite elements represent a homogeneous class (see Table 4). Being specified for a certain class, the pattern for C/EBP within composite elements cannot be used for a general search for C/EBP binding sites. Therefore, other patterns should be additionally provided for a more comprehensive search. Up to now we did not identify functional correlations for the other subgroups of the binding sites derived by the subtractive approach, neither in terms of the regulated genes nor of the C/EBP isoforms interacting with these sites (α,β, or others), although one of the consensi, C/EBP_alternative ("CATTKCSYCAKN"), resembles somewhat the half-site known for CREB/ATF-like bZIP factors (…CGTCA) known to heterodimerize with C/EBP (Shuman *et al.,* 1997). So the subtractive approach appears to be more general and does not depend on detailed biological knowledge.

It is interesting to analyze and compare the consensus sequences of the PWM obtained in the different subtraction cycles 2002 and 2005 with each other (see Table 4). All pairs of consensi are well aligned (inside the pairs); the first, the second and the fourth pairs share some similarities between the pairs. In spite of this, it would be hard to align them to get a common matrix of satisfactory quality. The differences between the first two consensi (of the first 2 rounds of subtraction) are minor, but they reflect the small deviations of the binding

patterns which may be crucial for the binding of C/EBP under different conditions.

As it has been mentioned in *Results* (and reflected in Table 4), the matrices found in the 3rd and the 4th rounds in 2002 and 2005 "mixed up" their places and correspond to each other "cross-wise". Nevertheless, the correspondence in the "mixed" pairs is good, and the reproducibility confirms the robustness of the approach. It seems that in the new training set the proportions of the subsets used in the last two rounds have been changed. The consensi giving the pair no.4 share some features with the motifs from the first two rounds of subtraction, while the consensi of the pair no.3 are definitely different. What may be the meaning of these findings? Of course, we run a risk of identifying something different, for instance a binding site for another TF, which may accompany a minority of the C/EBP sites. However, note that it was specifically the subset giving this consensus that has grown in the training set from 2002 to 2005, so that the corresponding motifs were found not in the 4th, as in 2002, but already in the 3rd round of subtraction. Probably that is not by chance. On the other hand, with some efforts we can try to find correspondence between the consensi of the 3rd pair and the others. The reverse complement of the 3rd consensus corresponds very nicely to other consensi in the last 4 positions:

```
      D G C A S A G G          (3rd consensus)
      C C T S T G C H          (reverse complement)
W N T G A T T G C T             (4th consensus)
        A T T K C Y T M A K     (2nd consensus)
          T T R C M C M A       (1st consensus)
```

Thus, it seems likely that this nucleotide pattern can mediate interaction with the transcription factor C/EBP under some conditions. To understand what makes this binding so specific and different from the other cases, we need more experimentally proven information. Hopefully, it will become available in the TRANSFAC database soon.

In spite of changes in the training set and the tools used in this work (Match™ and Gibbs Sampler), reproducibility of the results could be proven. This demonstrates the robustness of the approach.

As it has been mentioned in *Introduction*, PWM approach to TFBS prediction still keeps its first position among the other methods. In this work we do not try to substitute this method with something else, but try to refine one of its necessary steps, namely the construction of PWMs. To our knowledge, no one has tried so far to dissect the training sets prior to PWM construction.

Still, the reliability of single site predictions remains low. To be able to predict functional regulatory elements, we have to switch to the combinations of TFBS. In this work we deal mainly with TFBS pairs.

3.1.2. Distance distributions

3.1.2.1. The method: main idea, some methodological premise and the result

The main idea of this approach is that the distances between the TFBS in functional TFBS pairs have some specific distribution, which can be used as a distinguishing feature of functional pairs. As the existence of any preferred distance ranges is not self-evident (for instance, the distances could be distributed evenly), we had to investigate the behavior of real functional TFBS pairs and to see whether our assumption is correct (this will be discussed in 3.1.2.2).

A methodological novelty included in this approach is the theoretical modeling of random distributions, i.e., the distributions of distances between TFBS allocated randomly.

This is a question, which often raises before a researcher who deals with any subject of investigation: "What if this happens by chance?" Dealing with sequences of DNA regulatory regions and trying to understand the regularities of their structure, we repeatedly come across with this question. Roughly our interest in randomness can be expressed in three questions: Can this feature appear by chance in any sequence? Can it appear in any genomic sequence? Can it appear in any promoter?

The range of questions determines the choice of control sets normally used in promoter sequence analysis. To answer the first question, we use a set of random or randomized sequences (i.e., shuffled functional sequences). For the second question, the state of art is to take sequences of other genomic function, e.g. 2nd or 3rd exons (Pickert *et al.,* 1998). To tackle the last question, we can take promoters of another defined function or a set of randomly chosen promoters with any different function(s) (but not the one specific for the training set).

To use the random sets we have to select a reliable random number generator to be sure that the generated sequences are really random[1]. If the generator is sufficiently good, the next

[1] The quality of random number generators, included in many programing packages, varies, but even the best of them actually do not generate truly random numbers. A computer is a deterministic machine, and it is illogical to use a deterministic computer to generate sequences of random numbers (Gould, Tobochnik). However, it is possible to compute pseudorandom numbers which satisfy all statistical criteria for randomness. In such case the distinction between the truly random and pseudorandom numbers is unimportant. Nevertheless, referring to random number generators we mean, strictly speaking, pseudorandom number generators. The other problem is

problem is the estimation of the required number of random sequences in the set. This task is not trivial and, depending on the investigated feature, is sometimes hard to fulfil.

The exon sequences are usually well-defined and easy to collect. The problem is that these sequences, being coding sequences, have their own bias. The own characteristic features of the coding sequences will form a background which can interfere with the features of our interest. This can be especially harmful when we search, for example, for short over-represented motifs.

For the promoter sequence analysis, the control sets made of non-specific promoters are the most plausible. But the correct selection of the sequences for the set is not a trivial task for several reasons. First of all, we must mention that to collect a set of real (genuine, i.e. experimentally proven) promoters is problematic by itself. The best collection in the sense of reliability can be found in the Eukaryotic Promoter Database (EPD) (http://www.epd.isb-sib.ch) (Perier *et al*, 1999), but the number of the promoters is rather low (1871 for human, less for other vertebrate species). This amount would be enough for a control set, but if we want to make a sub-selection of promoters of genes having or, more important, lacking some specific function, the choice is very restricted. For some analyses we can use the whole set with the hope that the portion of genes in this set exerting the function that is specific for the promoters in our training set will be negligibly low. This is appropriate if we are sure that the collection of the promoters is random. This may not be the case for EPD since the manual annotation for that database may have been biased by the choice of the curators, or by the data available in literature.

Instead of the set of promoters we can use a set of sequences immediately upstream of TSS. Such sequences can be taken from DBTSS, a database for experimentally proven TSS (http://dbtss.hgc.jp/index.html). The quality of the sequences in the database is very good, since it gathers experimental data about genuine 5'-ends of cDNAs, but we have to remember that the location of a region cannot be taken as a guarantee of its functionality. Thus, the

---

that all the generators give a sequence which is repeated after some number of digits which is called period (Gould, Tobochnik). There are some kinds of problems requiring very long sequences for which the period of most of the usual packages is short, but for our class of problems the period length is satisfactory, unless we start to work with the whole genomes, etc.

The really random numbers can be generated from intrinsically random physical process, such as the time between clicks in a Geiger counter near a radioactive sample, or atmospheric noise. The problem of such sequences of random numbers is that the sequence is not reproducible; although we can principally store the outcome of a random physical process so that the random number sequences would be reproducible, such method would usually be inconvenient and inefficient for very long sequences (Gould, Tobochnik).

substitution of promoters with the regions around TSSs is a short-term arrangement, which we can apply until the pool of genuine promoters grows to satisfactory size. Thus, none of the discussed control sets is ideal. In the optimal case they should be all applied separately, with a consequent comparison of results, but this is too time-consuming, in particular when considering application of the methodology to high-throughput data.

Is there any alternative to the control sets? The alternative to an experiment is calculation, which is possible only when we know all properties of the investigated subject. Let us return to the questions I have listed above. There are three variants of sequences of our interest: regulatory, coding and random. Random sequences are the only ones the properties of which we know well enough to try to make some predictions.

In this part of work it has been shown that given the frequencies of TFBS, it is possible to theoretically model the distribution of distances in a random case.

Theoretically, the non-random occurrence of (real) TFBS in promoters may reveal itself in different ways. Given the fact that the set of the TFs working on each specific promoter is already not random, we must consider the other manifestations of non-randomness, such as the order of the sites, their orientation (mutual or to the transcription start site), the number of the sites of each type, and so on. It is conceivable that these are defined parameters as well which together govern the specific functionality of a set of promoters. The distance between the sites is only one of the characteristics, and concentrating on it, we should not underestimate the others. The positive side of this characteristic is that it is in principle possible to model it quite accurately.

We have modeled the distribution of distances between the constituents of heterologous TFBS pairs in random sequences. The random distance distribution depends only on the length of the sequence and the frequencies of the sites, which can be easily identified for each sequence (set of sequences) in every certain case. Now we have a tool to describe the randomness of the occurrence of TFBS pairs on certain distances, comparing the random distance distribution with the distributions in (sets of) sequences of interest. A difference from the random distribution by more than 3 standard deviations is considered as over- (or under-) representation. In this study we have considered only over-representation.

3.1.2.2. Application of the distance distribution approach

Developing this approach, we made a premise that in functional sequences there are some preferable distances, which are necessary for the functionality of the pair. To check the

plausibility of this assumption and to see how the new approach works, we applied it to the investigation of the behavior of real, experimentally proven composite elements. The functionality of the pairs of TFBS in CEs is already proven. Thus, we have only to see whether the behavior of the distances in CEs is in accordance with our expectations and whether we can use the distances as a distinguishing feature to detect potentially functional TFBS pairs.

We applied the method on the investigation of several sets of composite element-containing sequences (600bp centered around the composite element). The percentage of reidentification of the true positive TFBS pairs is high (86-100%) (Table 6). The filtering power of this approach for the selection of TFBS pairs for promoter models is also high: we can consider only the pairs in defined distance intervals instead of the whole stretch of several hundred base pairs, which shrinks the list of considered pairs by (roughly) two orders of magnitude, from several thousands to a couple of dozens (Fig. 11).

In some cases, we can also observe some peaks of over-represented distances, which do not contain true positive pairs. We could call them false positive, if we were sure in the absence of actual or potential activity of the sites forming these pairs. Unfortunately, this kind of information is normally unavailable. Thus, it is hard to judge about the role of the predicted pairs, found in the same sequences as the experimentally proven composite elements. The fact that they have not been reported as active sites does not mean that they can never play this role. We must remember that an experiment is not necessarily conducted under the same conditions that apply in reality (*in vivo*), so some relevant information can be easily missed. On the other hand, it may well be that the non-reported TFBS play a role of some "spare" sites, which can become active in case of damage of the "main" site(s). This is in agreement with the observation that the TFBS for some factors (NFAT, E2F) are enriched in promoter regions (Kel-Margoulis *et al.,* 2003).

All the additional peaks of over-represented distances were found at longer distances than the true positive ones. One possible explanation could be that in the sequences with a characterized short-distance pair the occurrence of long-distance additional pairs was not expected and, thus, experimentally checked.

Careful consideration of distance distributions can supply us with some kind of new information. For example, the distribution of distances between the NF-κB and Stat sites seems to reveal some periodicity (Fig. 11c). The centers of the peaks appear at practically equal distances of about 30bp (Table 6). This is accompanied by a high number of sequences possessing these pairs, and a relatively high number of these pairs in each sequence (on

average, each sequence of the NF-κB – Stat set has 2 to 4 predicted pairs of this type). This may be a hint on some interesting behavior of the binding sites for this pair of factors.

We find it remarkable that the known experimentally proven composite elements tend to appear exactly at statistically over-represented distances. This allows us to suggest that: (1) stretches of DNA necessary to bind a pair of cooperating TFs have discrete lengths; (2) there can be several discrete distances characterizing the same TFBS pair; (3) the length of a discrete stretch (set of them) is characteristic for a functional TFBS pair.

Our approach can be used as a filtering technique for the selection of TFBS pairs for promoter model construction. We successfully applied it to the construction of promoter models for LPS-triggered pathways as well as for the MyD88-dependent and –independent TLR-triggered pathways. The application of this method makes the process of the selection of candidate TFBS pairs much easier and quicker.

3.1.3. Other anti-false-positive measures

Three other approaches were devised to filter the numerous false positives which are unavoidable among the predictions of single TFBS. These approaches were then applied to the investigation of several examples of the defensive mechanisms of eukaryotic cells. One of the results of our work is a list of potential target genes triggered in the response human epithelial cell to the binding of *P.aeruginosa*; the list is enriched with different regulatory proteins, including transcription factors and known participants of the ascending pathways. This theoretical result must have two practical consequences: first, it allows to restrict further experimental research to a manageable number of candidate genes; second, it enables to understand or to clarify some uncertain details concerning the triggering pathways, and thus to make some new predictions based on this information. There is a number of published tools for searching for regulatory modules (i.e., “sequence elements that modulate transcription”, following the definition given by Bailey and Noble (2001) following Krivan and Wasserman,2001; Wasserman and Fickett,1998) (Hannenhalli S, Levy,2002; Frech *et al.,* 1997; Kondrakhin *et al.,* 1995; Prestridge, 1995; Berman *et al.,* 2002; Markstein *et al.,* 2002). The used algorithms may be divided in three classes (sliding window approach, hidden Markov models, discriminative technique), as briefly reviewed in (Bailey and Noble, 2001). Any of the approaches, independent of which algorithm it is based on, encounters the same problems arising from the biological nature (and extreme complexity) of the object: (i) scarcity of knowledge about exact location of promoters and enhancers and of experimentally proven binding sites (information used for constructing (+)-training sets); (ii) the fact that

statistical significance of a feature (TFBS or a cluster of them) does not necessarily tell anything about the biological functionality of this feature; analogously, the insignificance can not be taken as a proof of the lack of function; (iii) usually weak reasoning for grouping genes (their promoters) in sets according to their function, co-regulation, functional occurrence in the same cell types, etc. The latter has some lucky exceptions, like sets of muscle genes (Wasserman and Fickett, 1998) or cell-cycle regulated genes (Kel *et al.,* 2001), and the situation will obviously improve with further development of microarray technique.

In the present work we tried to address the listed problems. We could not, of course, improve the situation with the paucity of experimental data, only endeavored to make our data searches as accurate and exhaustive as possible. In principle we developed our approaches basing them, whenever possible, on biological reasoning. We find it extremely important to use as much experimental evidence as it is available at the moment. In our approach we alternated two different kinds of steps - expanding the data and restricting it: exhaustive data search – "seed" and distance constraints – exhaustive enumeration of all possible pairs – complementary pair constraints.

To avoid the problem of low confidence in the (+)-training set (which may occur not only in our specific case), we developed the approach of "seed" sequences. The difference from the "seeds" used in cluster analysis is that in our approach the choice of the "seed" is biologically based. Although the "seed" approach is, obviously, a restrictive measure, moreover, a pre-process restriction, which may result in missing potentially relevant additional sequence features, we find it useful and appropriate when the choice of the "seed" is made on a solid biological basis. After having applied the restrictive "seed" technique and distance assumptions, we undertake an exhaustive, complete enumeration of all possible pairs of potential TF binding sites that can be found in the (+)-training set, which in turn reveals a large number of combinations. This list of all found pairs is processed under a new kind of constraints imposed by the search for complementary pairs.

The search for complementary pairs is a completely new approach, which supplies us with a new kind of information. It enables to identify subsets of the (+)-training set, which possess different regulatory modules, thus suggesting their triggering by different regulatory pathways. This kind of information becomes extremely important in two cases: (i) when two or more pathways are presupposed to be triggered in the cellular response, like in the case considered in this work; (ii) when the (+)-training set consists of not really co-regulated, but of co-expressed genes, without precise information about which of them are regulated by the same mechanism. The identification of complementary pairs and, consequently, groups of

sequences enables to better define the co-regulated genes thus providing a partial, although only predicted, confirmation of the co-regulation, and at the same time to understand better the ascending pathways.

The final result of our search supported the idea of complementary pairs. We could re-identify the LPS-responsive subset in the pool of the *P.aeruginosa*-triggered genes, which was in a good agreement with the experimental data, and in the other part of work we could distinguish between the MyD88-dependent and –independent pathway with the help of the complementary pair approach. We will return to these examples in the discussion of applications.

3.2. Applications

The examples of applications of two of the approaches developed in this work – the subtractive approach to PWM generation and the method of distance distributions – have already been described in the corresponding sections. There these approaches were applied to special cases, and the purposes of the applications were rather confined. The first approach stands a bit apart from the others, because we do not need to apply it in any promoter model construction. It is more a preventive measure used to obtain better matrices in those cases where only very weak patterns are available so far; these improved matrices will then be used for the promoter models. Thus, having once applied it to the improvement of the PWM for a certain kind of transcription factor (C/EBP) which we needed for our models, we did not return to this approach in the next applications. As for the second approach, the method of distance distributions, its application to the investigation of composite elements was mainly necessary to confirm the initial premise that the distance distributions specifying every functional TFBS pair exhibit typical characteristics. After this had been demonstrated, we applied this method to the promoter model construction together with the other approaches.

This part of the Discussion is devoted to the applications of all developed approaches, so that they form a kind of a small “pipeline”.

The first promoter model, called “Pa-model”, was constructed for the response of human epithelial cells to the interaction with *P. aeruginosa*. This is the only model constructed without the application of the approach of distance distributions (which was devised two years later). Being curious whether we can see the pairs included in the promoter model with the new approach, we applied the method of distance distributions and could re-identify practically all pairs of the model as pairs at over-represented distances (see *Appendix 3c*). This

proves not only the robustness of the results, but demonstrates once more the usefulness of the distance distribution approach. The advantage of the approach is that its application makes the search very quick and efficient.

The "P.a.-model" was applied to screening of 13,000 upstream regions of human genes and identified 430 new target genes, which are potentially involved in antibacterial defense mechanisms. We realize that 430 is nearly exactly what we can expect as the number of false positives when starting from a total of 13000 human promoters (the rate of false positives was 3.4%). On the other hand, we realize also that the number of true genes which we can expect to be triggered during this response can be lower or of the same order of magnitude (see, for example, Eskra *et al.,* 2003; Bandman *et al.,* 2002; Krappmann *et al.,* 2004; Ichikawa *et al.,* 2000). We believe that the true positive genes are among our predictions, but it is hard to distinguish them on the background of the false positives. Thus, the statistically reasonable false positive rate (3.4% against 52% of true positives) still appears to be too high for practical applications. The false positive rate which should be achieved by a predictive promoter model should be much less (in optimal case less than one order of magnitude) than the expected number of true positives. We cannot suggest a method for such an estimation, but if we, simply based on empirical evidence, very roughly estimate the number of genes expressed in one round of a triggered response as several dozens to couple of hundreds, the desired level of the FP rate can be estimated as 0.3%.

Does this mean that the proposed model and the predictions made with it are senseless? Of course, not. One of the first aims of the construction of a promoter model and prediction of potential target genes is to generate hypotheses for more efficient further experimental verification. Our predictions restrict the number of genes to be experimentally checked for plausibility to only approximately 3% of all genes, which should make the life of experimentalists much easier. On the other hand, to predict the potential target genes is not the only purpose of a promoter model. We have already discussed above (in 3.1.3) how the identification of complementary pairs can help the understanding of the correspondent pathways. Here I would like to return to this subject and to show how the approach worked in the concrete case of the *P. aeruginosa*-triggering.

There is a lot of evidence in literature that interleukin 8, β-defensin, monocyte chemoattractant protein and different mucins are regulated through LPS-triggered pathway(s) (Harder *et al.,* 2000 ; Diamond *et al.,*2000 ; Diamond *et al.,*1993 ; Singh *et al.,*1998 ; Liu *et al.,* 1998; Ratner *et al.,* 2001; Ko *et al.,* 1997 ; Sar *et al.,* 2000 Smith *et al.,* 2001 ; Zhang and Ghosh 2001; Leidal *et al.,* 2001 ; Sar *et al.,* 1999 ; Mori *et al.,* 1999; Li *et al.,* 1998; Kovarik

et al., 1996; Perrais *et al.,* 2001; Walsh *et al.,* 2001; Gum *et al.,* 1997). On the other hand, it is also well-known that LPS is one of the major triggers of the antibacterial response (see *Introduction*, 1.2.3.). We know that in the particular case of interaction with *P. aeruginosa* this pathway is not the only one (Imundo *et al.,* 1995; Prince, 1992; Sheth *et al.,* 1994), but we do not know in advance which of the genes in the (+)-training set belongs to which pathway (except for several genes as listed above). We had no means to include our pre-knowledge in the search. With the complementary pair approach we could re-identify the LPS subset in good agreement with our expectations (Table 9). This not only confirms the efficiency of the method, but allows to assign some other genes, like ELF3 and cytochrome p450, to the same pathway; moreover, we get another, although indirect, evidence of the cooperation of the transcription factors included in the model, which (again indirectly) confirms their belonging to the LPS-triggered signaling pathway.

The situation with the complementary pairs in the next example (of the MyD88-dependent and –independent pathways) is different from the case of *P.aeruginosa*-triggering. In this case the set was specially organized in such a way, that we knew in advance which genes belonged to which pathway. This allowed us to check the method once more and to see which combinations of transcription factors work in every particular pathway. The distinction between the complementary pathways was very clear. All combinations of the MyD88-independent pathway contained IRF TFBS. Only 3 sequences from 48 in the MALP-dependent subset were found to contain the NF-κB-IRF composite element. A check in the TRANSFAC database revealed that one of these additional sequences (RANTES) possesses a known composite element of this type. The two other sequences sharing the NF-κB - IRF pattern (iNOS and IκBα) have not been reported as IRF-dependent and are well-known MALP- as well as LPS-triggered genes. Both proteins are key participants of many pathways and cannot be restricted to only the MALP-2 signaling. Thus, they may well belong to the MyD88-independent pathway as well.

The combinations found to characterize the MALP-dependent subset revealed some especially interesting information. One of the pairs contains CREB and SRF, which are both substrates for RSK2 (according to the scheme in Fig. 4). The confirmed special role of CREB is more interesting in connection with simultaneous phosphorylation of histone H3, which has been shown to be phosphorylated together with CREB by RSK2 in response to mitogenic stimulation by epidermal growth factor (Merienne *et al,* 2001). Confirmations of the dependency of CREB and H3 phosphorylation on RSK has been shown or mentioned in some other papers (Stevenson *et al,* 1999, Vaidyanathan and Ramos, 2003). The statement that

RSK2 has a critical role as an effector of RAS-mitogen-activated pathway and regulator of immediate response, made in one of these articles (Stevenson *et al,* 1999), gets a new, although indirect, confirmation through the promoter model developed here: both CREB and SRF point at RSK2. The indirect hint on the involvement of H3 phosphorylation in MALP signaling pathway given by the presence of CREB in the promoter model would deserve special experimental investigation. This is a good example of how a promoter model can help in understanding the regulatory network.

CREB, but not SRF, is also involved in the promoter models for LPS triggering. The absence of SRF might mean that the TFs are activated by other pathways, but we would not like to draw premature conclusions. The case of LPS triggering was the only one where we could not make a final decision of the most appropriate promoter model and suggested three models giving approximately same results. The reason for such uncertainty lies in the quality of the training set. Selecting the training sets, we depend completely on the experimental data. As it has been discussed above (in paragraph 2.3.1), every experimental method has its rate of false positives, which results in some error in our true positive sets. Not knowing exactly how many genes can be expected to possess the specific combination, we have to reduce significantly the rate of the expected true positives; for instance, in the case of LPS model we accepted pairs, which were present in not less than 65% of the TP set. The reduction of the true positive rate has the obvious consequence of yielding a higher rate of false positives. The only way out in this context is the improvement of the quality of true positive sets. In the case of the LPS-triggered genes we are very optimistic, because we collaborate with the experimentally working group of C. Scheidereit (MDC Berlin). Presently we are looking forward for the work with new, better defined sets, which will allow to refine our models.

3.3. Shortcomings

Our approach, as any other, has its limits. It has been shown for the genuine composite elements of certain types (for instance, NF-AT and AP-1) (Kel *et al.,* 1999) that one of the two constituents of a composite element could be rather degenerate, as compared with its canonical consensus sequence or when scored with a positional weight matrix (PWM). This means, that our requirement for all binding sites to be found with rather high PWM thresholds may be too restrictive. We are running risk to overlook those constituents of pairs, which possess weak consensi. It is difficult to find a solution to this problem. We have no information about which of the TFs could be represented by such low-threshold consensus, and if we take from the very beginning the lower thresholds for all considered matrices, we

will be drowned in potential binding sites, nearly all of them probably being false positives. Nevertheless, it is known that the PWM approach is better than string identification, which even with allowed mismatches cannot provide the same flexibility as PWMs (Quandt *et al.,* 1995).

The next source of limitations we see in the preselection of factors according to published data. Obviously, we can not expect that the experimental data is exhaustive; some of the transcription factors may not yet be reported just because their participation in a certain process has not yet been investigated. On the other hand, statistical over-representation, as it has already been mentioned before, can not be taken by itself as proof of biological functionality or its lack; some TFBS may not be found over-represented due to their degenerate nature, when the structure of TFBS allows it to appear practically in any sequence, being functional or not, so the occurrence cannot be used as a criterion of functionality. We had no other idea of how to take into account those TFBS which are not over-represented, but to rely on published experimental data. We find that the usual methods based on statistical over-representation are even more restrictive, but maybe the best solution could be found in merging both approaches – i.e., using the experimental evidence along with statistical ones, for instance using Bayesian techniques.

The shortcomings connected with the quality of experimental data used as an input in our approaches has been intensively discussed in the previous sections. Actually, these are not the drawbacks of the bioinformatics approach, but more the consequences of the quality of the experiments themselves, which influence our results. This situation, of course, will improve with the further development of the experimental techniques and with the mere accumulation of experimental data, because the cross analysis of even contradictory data, with the analysis of discrepancies and coincidence, can supply with interesting information.

3.4. Related work

Among the available tools dealing with sequence analysis (Tables 2 and 3) only several concentrate their efforts on prediction of combinations (clusters) of motifs of TFBS (e.g., CisMols Analyser, Cister, COMET, FrameWorker (combined with other tools of Genomatix), MSCAN). The other tools deal with motif or TFBS prediction based on statistical over-representation (e.g., Ann-spec, POCO, POBO, BioProspector, Mdscan) or additional incorporation of comparative genomics features (e.g., ConSite, FOOTER, CompareProspector). In our approaches we used GIBBS Sampler for motif search and Match tool for TFBS prediction. Match was selected since it uses the library of TRANSFAC

matrices for TFBS, which is today the richest collection of PWMs. The identification of motifs or potential TFBS is a necessary step, but it is only one of several steps of promoter modeling; the more comprehensive view is suggested only by Genomatix Suite. We would like to note that the set of methods and the logics underlying them appear to be the closest to our approaches, as far as we could judge since Genomatix Suite is a commercial package.

Several steps of promoter modeling are made in Genomatix by consecutive application of the next tools: BiblioSphere™, responsible for literature mining; ElDorado™, undertaking genome annotation; GEMS Launcher, which includes: MatInspector searching for TFBS with PWMs, ModelInspector searching for complex regulatory patterns and FrameWorker, automatically defining of a common framework of TFBS. The applications of these tools are described in several papers (Kramer-Hammerle *et al.,* 2005; Seifert *et al.,* 2005; Tasheva *et al.,* 2004).

We have compared the approaches applied in these papers with ours. The approaches to TFBS prediction (made by MatInspector and Match) are very similar; accordingly, they give very similar results. The extensive literature search undertaken with the help of the BiblioSphere tool resembles our approach of "seed" sets. The important difference is that after the search for pairs in a "seed" set we extrapolate the obtained results to the rest of the training set, thus never completely neglecting the rest of the true positive information, whereas in the Genomatix applications the authors do not return to the initial data and make all predictions based only on the results obtained for a very small subset of the training set. The work with less than 10% of co-expressed genes and even a successful identification of a set of characteristic features of this tiny subset hardly can give an explanation why the other 90% of the genes of the microarray set revealed any overexpression. In the paper of Kramer-Hammerle *et al.* (2005) the authors are content with the model re-identifying ~1% of the investigated set, - the level which would be out of interest in our investigations. The authors do not mention the rate of false positives gained by their method, but we are ready to admit that it should be very low; the price for that is the obvious very low sensitivity, but we can assume that the high sensitivity was not the aim of these investigations. In our approach, the "seed" set has normally the size of ~30% of the training set, and the lowest acceptable number of re-identified true positives is 50%. As a result, our approach gives more comprehensive information regarding the initial set of co-expressed genes.

Apparently, the aims of investigating expression data in our work and in the papers listed above are different. We tried to answer the question: "Which features allowed the genes to be co-expressed?", whereas they tried to find a specific subset in the whole set with a defined

function, to identify the characteristics of only these several genes and to find their place in known interaction networks.

Only very few approaches developed so far take into account the distance between the binding sites as a discriminative characteristic of a regulatory module. To our knowledge, there is only one publication in which the authors developed a method for discrimination of a specific distance between the binding sites from a distance which can occur in a random sequence (Makeev *et al.,* 2003). They considered the distances only between neighbouring binding sites, not taking into account all possible pairs as it has been done in our approach. Thus, our approach can be regarded as a generalization of their method. The other important innovation of our approach is the allowance of the shift of binding sites in the pair, when we consider the pairs on some distance plus an interval δ. We did not observe the periodic signals in all investigated cases (only in several few), as V. Makeev and colleagues did, but we think this deserves further investigation.

The approach of complementary pairs and subtractive approach of matrix generation are novel and we can hardly find any analogous tools provided by other groups.

3.5. Perspectives

We see the perspectives of this work in two different fields: further investigation of regulatory networks and further development of the methodological approaches, making them more flexible and applicable to any similar task. The list of predicted target genes has to be evaluated experimentally, but may have its value for further research already on the present step. The future work on reconstructing the intracellular pathways triggering the genetic program of the antibacterial cell response will be well supported with the information picked up from this list. It may give some hints for the next steps of experimental research, for instance providing information about the first candidates to be checked. The information about the complementary subsets of regulated genes helps to better understand the triggering pathways, and the complementarity of their function is a subject for further consideration.

The obtained promoter models, especially the complementary combinations, point at some specific pathways, which participate in their triggering. This information has to be verified experimentally. We find it especially interesting to get experimental confirmation of the importance of the RSK2-dependent pathway in MALP-2-triggering.

Some systems that were under our consideration require more work. In the case of LPS-triggered genes we do not consider the work finished, and are expecting the prolongation of

these investigations together with the experimentalists of the C.Scheidereit group (MDC Berlin).

The methodological approaches presented in this work can, of course, be applied to other objects. In this work we focused on the experimentally proven basis for the initial choice of transcription factors. This kind of evidence is stronger than any prediction, but it can work only when this information is available, which may be not the case for some other sets of genes or cellular situations. In the next step of development we would like to allow also an exhaustive computational search through the whole list of known TFs for potential constituents of the models. The usage of Bayesian techniques, as mentioned in the previous paragraph, would be also appropriate for this kind of predictions.

The proposed methods are intended to become a part of a pipeline XAP (eXpression Analysis Pipeline/Platform) which is under creation in our department. The system starts with the analysis of microarray data and proceeds with the biological interpretation of the identified gene set, which includes the promoter modeling.

4. MATERIALS AND METHODS

4.1. Software

Mathcad®, version 11 (Mathcad 2001*i*). Mathcad® is a comprehensive design environment that enables to explore, calculate and document mathematical formulas, methods and values during the design phase of a product. Using standard math notation, Mathcad enables to combine formulas, text and interactive graphics in a single worksheet.
All the software for TFBS pairs' identification, distance distributions and other used in this work has been written in Mathcad.

4.2. Tools

- **Match^{TM}** (http://www.biobase.de/cgi-bin/biobase/transfac/start.cgi) (Kel *et al.,* 2003).

The Match^{TM} tool is designed for searching potential binding sites for transcription factors in any sequence which may be of interest. Match^{TM} uses a library of mononucleotide weight matrices from TRANSFAC®. Match was used to predict potential TFBS (see 4.6).

- **Gibbs Motif Sampler** (http://bayesweb.wadsworth.org/gibbs/gibbs.html) (Thompson *et al.,* 2003).

The Gibbs Motif Sampler is a software package for locating common elements in collections of biopolymer sequences. Gibbs Sampler was used for identification of common motifs in sets of TFBS-containing sequences for subtractive approach to matrix generation.

4.3. Databases

- **Eukaryotic Promoter Database** (http://www.epd.isb-sib.ch), release 77-1 (Perier *et al.,* 1999). The Eukaryotic Promoter Database is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally.
- **Ensembl Genome Browser** (http://www.ensembl.org/index.html) (Birney *et al.,* 2004).

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

- **DBTSS**, the database of transcription start sites (http://dbtss.hgc.jp/index.html), release 3.0 (Suzuki *et al.,* 2002, 2004). Based on human and mouse full-length cDNA sequences,

DBTSS contains exact information of the genomic positions of the transcriptional start sites and the adjacent promoters for 8,793 and 6,875 human and mouse genes, respectively (status for October 2005, release 4.0). Of these, 3,324 can be paired as mutually homologous genes between human and mouse and their promoters can be compared with each other.

- **TRANSCompel®** Professional release 6.1, 8.4 (http://www.biobase.de) (Kel-Margoulis *et al.,* 2002). TRANSCompel is a specialized manually curated database on composite regulatory elements; it is an extension module to the TRANSFAC database system. Presently it contains information about 422 composite elements based on 1458 evidence extracted from 497 references (status for October 2005, release 9.2)
- **TRANSFAC®** Professional release 6.1, 9.1 (http://www.biobase.de) (E.Wingender *et al.,* 2000). TRANSFAC is a manually curated database on transcription factors; it presents currently the largest archive of transcription factors, their binding sites and a unique library of positional weight matrices. It contains the information about in total 7520 transcription factors, 15643 binding sites and 762 PWM (status for October 2005, release 9.2).
- **TRANSPATH®** Professional release 4.1 (http://www.biobase.de) (Schacherer *et al.,* 2001). TRANSPATH is an information system on gene-regulatory pathways, and an extension module to the TRANSFAC database system. It focuses on pathways involved in the regulation of transcription factors in different species. The states of elements of the relevant signal transduction pathways (such as complexes, signaling molecules) are stored together with information about their interaction in an object-oriented database.
- **TRANSPRO®** Professional release 2.1 (http://www.biobase.de). It contains upstream (5') sequences of human, mouse, and rat genes, together with extensive annotation. Presently TRANSPRO contains sequences for 13396 human, 14605 mouse and 22953 rat promoters (release 2.2, June 2005).

4.4. Training sequence sets

4.4.1. Positive training sets

- **(+)-Training set for the *P. aeruginosa*-response study**

Positive (+) training set comprises:

2. Promoters of human genes shown to be expressed in epithelial cells after interaction with *P. aeruginosa* by means of:
    a. microarray analysis (Ichikawa *et al.,* 2000),
    b. other methods (Smith *et al.,* 2001; McNamara *et al.,* 2001; Harder *et al.,* 2000;

Leidal *et al.,* 2001; Kovarik *et al.,* 1996; Walsh *et al.,* 2001; Becker *et al.,* 2001).(Table 15)

3. Orthologous mouse promoters.

The sequences were derived either from Eukaryotic Promoter Database (http://www.epd.isb-sib.ch), or from DBTSS, the database of proven transcription start sites (http://dbtss.hgc.jp/index.html). The length of the sequences was 600 bp (-500/+100). This region comprises most of the known upstream elements and corresponds to the upstream region used by Davuluri *et al*. as "proximal promoters" for promoter recognition (Davuluri *et al.,* 2001), plus a 100 bp proximal downstream region which also contains many known regulatory elements documented in the TRANSFAC database (Matys *et al.,* 2003).

The "seed" set is a subset of the positive training set manually selected for highest experimental reliability (see Table 15). The whole set is available in "Supplementary materials/Pa_33seq_600bp.doc").

- **(+)-Training set for the LPS-responsive genes**

The set was based on the data published by D.Krappmann *et al* (2004) (Table 16). The upstream sequences for the genes triggered after 90 min of LPS treatment were derived either from Eukaryotic Promoter Database (EPD), or from DBTSS. Only the genes, which performed more than 2-fold overexpression, were chosen for the set. The length of the sequences was 600 bp (-500/+100). The set is available in "Supplementary materials/91seq_LPS.doc").

- **The set of C/EBP TFBS-containing sequences used for refining C/EBP matrix by subtractive approach**

The set consisted of sequences containing experimentally proven transcription factor binding sites for C/EBP taken from the TRANSFAC® database. Every binding site was prolonged by 8 nucleotides to the either side. In 2002 (TRANSFAC release 6.1) the set comprised 164 binding site entries for C/EBP in non-artificial sequences; in 2005 (TRANSFAC release 9.1) the set comprised 193 sequences (see "Supplementary materials/subtractive approach/Subtractive_seq_sets.doc", S1 and S2, respectively).

- **Composite element (CE)-containing sequences**

Eight sets comprising the most numerous experimentally proven composite elements described in the TRANSCompel® Professional database, release 8.4, (http://www.biobase.de) (Kel-Margoulis *et al.*, 2002) were chosen as positive training sets: AP-1–NF-κB (13

sequences), AP-1–ETS (15 seq.), AP-1–NFAT (10 seq.), NF-κB–C/EBP (14 seq.), NF-κB–IRF (9 seq.), NF-κB–Stat (7 seq.), NF-κB–HMG I(Y) (7 seq.), IRF–PU.1 (7 seq.) The sequences containing the corresponding CE were prolonged around the reported binding sites by 300 bp to either side from the center of the CE. The sets are available in "Supplementary materials/CE-containing seq/XX.doc")

Table 15. The genes of the (+)-training set (without orthologs) for the *P. aeruginosa*-response study. Marked with asterisks are those included in the "seed" set.

No	Gene name	Accession no. And LocusLink	Experimental evidence	Additional information	Participation in anti-Pseudomonas response
1	Monocyte chemoattractant protein-1, MCP-1*	EMBL: D26087	Microarray (Ichikawa *et al.,* 2000), other experiments (Ratner *et al.,* 2001; Ko *et al.,* 1997 ; Sar *et al.,* 2000)	Is well know as expressed in antibacterial response	proven
2	β –defensin*	LocusLink: 1673	Harder *et al.,* 2000 ; Diamond *et al.,* 2000 ; Diamond *et al.,* 1993 ; Singh *et al.,* 1998 ; Liu *et al.,* 1998	Is well known as expressed in antibacterial response; important target gene in innate immunity	proven
3	Interferon regulatory factor 1, IRF-1*	LocusLink:3659	Microarray (Ichikawa *et al.,* 2000)	Known to be expressed in epithelial cells	probable
4	Equilibrate nucleoside transporter 1, SLC29a1	LocusLink: 2030	Microarray (Ichikawa *et al.,* 2000)		
5	Proteinkinase C η type, PKCη*	LocusLink: 5583	Microarray (Ichikawa *et al.,* 2000),TRANSPATH®	Important link in Ca^{2+}-connected pathways	probable
6	Folypolyglutamate synthase, FPGS	Ensembl: ENSG00000136877	Microarray (Ichikawa *et al.,* 2000)		
7	RhoB*	LocusLink: 388	Microarray (Ichikawa *et al.,* 2000)	is induced as part of the immediate early response in different systems	probable
8	Origin recognition complex subunit 2, hORC2L	LocusLink: 4999	Microarray (Ichikawa *et al.,* 2000)		
9	Transcription factor TEL2*	LocusLink: 51513	Microarray (Ichikawa *et al.,* 2000)	Transcription factor	probable
10	Interleukin 8, IL8*	EPD: EP73083 LocusLink: 3576	Smith *et al.,* 2001 ; Zhang and Ghosh 2001; Leidal *et al.,* 2001 ; Sar *et al.,* 1999 ; Mori *et al.,* 1999	Is well know as expressed in antibacterial response	proven
11	Transcription factor ELF3*	LocusLink: 1999	Microarray (Ichikawa *et al.,* 2000)	Transcription factor	probable
12	Mucin 1(mouse gene), MUC1*	RefSeq: NM_013605	Li *et al.,* 1998; Kovarik *et al.,* 1996; Perrais *et al.,* 2001; Walsh *et al.,* 2001; Gum *et al.,* 1997	Different mucins are shown as expressed in antibacterial response	proven
13	NF-κB inhibitor α, IκBα*	LocusLink: 4792 EPD: EP73215	Microarray (Ichikawa *et al.,* 2000)	the main link in NF-κB-targeting pathways	Highly probable
14	Tissue Factor Pathway Inhibitor 2 ,TFPI	LocusLink: 7980 EPD: EP73430	Microarray (Ichikawa *et al.,* 2000)		
15	Urokinase-type plasminogen activator precursor, PLAU	LocusLink: 5328	Microarray (Ichikawa *et al.,* 2000)		
16	c-jun*	LocusLink :	Microarray (Ichikawa *et al.,* 2000)	Transcription factor	probable
17	Cytochrom P450 dioxin-inducible*	LocusLink: 1545	Microarray (Ichikawa *et al.,* 2000)]	Stress-inducible	probable
18	Dyphtheria toxin resistance protein, DPH2L2	EPD: EP74285	Microarray (Ichikawa *et al.,* 2000)]		

Table 16. Genes selected for the true positive set representing early LPS-triggered genes. Locus links are given for *Mus musculus*; for every mouse sequence a human ortolog was also considered.

Gene name	LocusLink	Gene name	LocusLink
TNF-β, tumor necrosis factor beta	LocusLink: 16992	PKCδ, protein kinase C, δ type	LocusLink: 18753
TNF-α, tumor necrosis factor alpha	LocusLink: 21926	ABIN-1, A20-binding inhibitor of NF-κB activation	LocusLink: 57783
IκBα, nuclear factor of κ light polypeptide gene enhancer in B-cells inhibitor, α	LocusLink: 18035	IAP-1, baculoviral IAP repeat-containing protein 3	LocusLink: 11796
MIP-1α, macrophage inflammatory protein 1, alpha	LocusLink: 20302	BFL-1, BCL-2-related protein A1	LocusLink: 12047, 12044, 12045
4-1BBL, tumor necrosis factor ligand family member 9	LocusLink: 21950	JUN-B, transcription factor	LocusLink: 16477
IP-10/CXCL10, small inducible cytokine B10 precursor	LocusLink: 15945	MAIL, molecule possessing ankyrin-repeats induced by lypopolysaccharideE	LocusLink: 80859
tumor necrosis factor C	LocusLink: 16994	RELB, transcription factor	LocusLink: 19698
WNT10A, WNT-10A protein precursor	LocusLink: 22409	ERF-1(TIS11B), butyrate response factor 1	LocusLink: 12192
EBI-3, Estein-Barr virus induced gene 3	LocusLink: 50498	MPC2 chomobox protein homolog 4	LocusLink: 12418
RANK, receptor activator of nuclear factor κB ligand	LocusLink: 21943	NAB2 NGFI-A binding protein 2	LocusLink: 17937
IL-12, interleukin 12 alpha chain	LocusLink: 16159	IKB-ε NF-κB inhibitor ε	LocusLink: 18037
ICAM-1, intracellular adhesion molecule-1	LocusLink: 15894	OCT2 octamer-binding transcription factor 2	LocusLink: 18987
CD40, tumor necrosis factor receptor family member 5	LocusLink: 21939	B-MYB, MYB-related protein B	LocusLink: 17865
NOTCH 1, neurogenic locus NOTCH homolog protein 1	LocusLink: 18128	ICSBP, interferon consensus sequence binding protein	LocusLink: 15900
Eph receptorA2, Ephrin type-A receptor 2	LocusLink: 13836	Stra13, stimulated with retinoic acid 13	LocusLink: 20893
IL-10R-B, interleukin 10 receptor β chain	LocusLink: 16155	CHOP10, C/EBP-homologous protein	LocusLink: 13198
CCR-7/ EBI1 C-C chemokine receptor type 7	LocusLink: 12775	IRF-5, interferon regulatory factor 5	LocusLink: 27056
CD5/Ly-1, T-cell surface glycoprotein CD5	LocusLink: 12507	BTG2, NGF-inducible protein TIS21	LocusLink: 12227
A20, putative DNA binding protein	LocusLink: 21929	p105, nuclear factor NF-κB P105 subunit	LocusLink: 18033
Map3k8, mitogen-activated protein kinase kinase kinase 8	LocusLink: 26410	NUR77, nuclear hormone receptor	LocusLink: 15370
MYD118, myeloid differentiation primary response protein	LocusLink: 17873	NF-ATC1, nuclear factor of activated T-cells, cytoplasmic 1	LocusLink: 18018
PEA-15, asrtocytic phosphoprotein 15	LocusLink: 18611	TIS7, interferon –related developmental regulator 1	LocusLink: 15982
PAC-1, dual specificity protein phosphatase 2	LocusLink: 13537	FIG-1(IL4ind1), interleukin-4 induced protein 1	LocusLink: 14204
CASPASE-11	LocusLink: 12363	ZNF151, zink finger protein 151	LocusLink: 22642
TRAF3, TNF receptor associated factor 3	LocusLink: 22031	SWAP70, SWAP complex protein, 70 kDa	LocusLink: 20947
BCL-10, B cell lymphome/leukemia 10	LocusLink: 12042	RGS16, regulator of G-protein signaling 16	LocusLink: 19734
F52/MLP1, macrophage myristoylated alanine-rich C kinase substrate (brain protein F52)	LocusLink: 17357		

- **(+)-Training set for the genes triggered through MyD88-dependent and – independent pathways.**

MyD88-dependent genes were represented by the set of MALP-2-triggered genes. The subset consisted of promoter (or 5'-upstream) sequences of MALP-triggered genes, collected basing on published experimental data (see table 17A). MyD88-independent TLR-4-triggered genes were represented by a set of IRF-responsive genes. The subset contained promoter or 5'-upstream sequences of IRF-triggered genes, as reported in literature and TRANSFAC® (see Table 17B).

The sequences were taken from EPD or DBTSS. The length of the sequences was 600 bp (-500/+100). The set is available in "Supplementary materials/58seq_MALP-IRF.doc").

4.4.2. Negative training (Control) set

The negative (-) training set was composed of randomly chosen 5'-upstream sequences derived from the TRANSGENOME information resource of annotated human genome features (Kel-Margoulis *et al.,* 2003). The set was manually cleaned from all genes, which potentially could be involved in the same or similar cellular responses. The set comprised 2067 sequences (see "Supplementary materials/Control_2067.doc").

Table 17. MALP-2-triggered and IRF-responsive genes used for the construction of positive training set.

<table>
<tr><td colspan="4">A. MALP-2-induced genes</td></tr>
<tr><td>Gene name</td><td>reference</td><td>Gene name</td><td>reference</td></tr>
<tr><td>CD80</td><td rowspan="3">Rharbaoui et al., 2002;
Weigt et al., 2003;
Link et al., 2004</td><td>IL8</td><td rowspan="15">Deiters et al., 2004</td></tr>
<tr><td>CD83</td><td>MIP-1</td></tr>
<tr><td>CD86</td><td>MIP-2a</td></tr>
<tr><td>iNOS</td><td>Muhlradt et al., 1997</td><td>Cathepsin h</td></tr>
<tr><td>c-fos</td><td>Quentmeier et al., 1994</td><td>G-CSF</td></tr>
<tr><td>COX2</td><td>Muhlradt and Schade, 1991</td><td>LIF</td></tr>
<tr><td>TNF-α</td><td rowspan="3">Muhlradt and Schade, 1991;
Deiters et al., 2004</td><td>MMP-11</td></tr>
<tr><td>IL-6</td><td>PDGF-B</td></tr>
<tr><td>IL-1β</td><td>PLGF</td></tr>
<tr><td>VEGF A</td><td rowspan="3">Deiters et al., 2004</td><td>TIMP-1</td></tr>
<tr><td>GM-CSF</td><td>RANTES</td></tr>
<tr><td>MCP-1</td><td>PDGF-A</td></tr>
<tr><td colspan="4">B. IRF-responsive genes</td></tr>
<tr><td>Gene name</td><td colspan="3">reference</td></tr>
<tr><td>IP-10/CXCL10</td><td colspan="3">Kawai et al., 2001; Sato et al., 2002; Ohmori and Hamilton, 1993</td></tr>
<tr><td>IFNβ</td><td colspan="3">Schafer et al.,1998; Watanabe et al., 1991; TRANSFAC®</td></tr>
<tr><td>IFNα</td><td colspan="3">Sato et al., 2002; Barnes et al., 2001; TRANSFAC®</td></tr>
<tr><td>ISG 15K</td><td colspan="3">Grandvaux et al., 2002; TRANSFAC®</td></tr>
<tr><td>GARG16</td><td colspan="3">Kawai et al., 2001; Sato et al., 2002</td></tr>
<tr><td>ISG 54K</td><td colspan="3">Grandvaux et al., 2002; Navaro et al., 1998; TRANSFAC®</td></tr>
<tr><td>IL15</td><td colspan="3">Azimi et al., 2000; TRANSFAC®; TRANSCompel®</td></tr>
</table>

4.5. Defining the sets of transcription factors (potential constituents of the model)

We based our selection of TFs on experimental evidence. For that we undertook an extended literature search, looking for the TFs which have been shown to take part either directly in the response for which the model was made (e.g., of epithelial cells to *P. aeruginosa* binding) or in the pathways triggered during similar responses.

4.5.1. Model for *P.aeruginosa* triggering

The search revealed 5 candidate factors: NF-κB (Bergmann *et al.,* 1998; Guha and Mackman, 2001; Harder *et al.,* 2000; Ko *et al.,* 1997; Li *et al.,* 1998; Smith *et al.,* 2001; Voynow *et al.,* 1999; Zhang and Ghosh, 2001) C/EBP (Ben-Baruch *et al.,* 1995; Guha and Mackman, 2001; Ko *et al.,* 1997), AP-1 (Ben-Baruch *et al.,* 1995; Guha and Mackman, 2001), Elk-1 (Guha and Mackman, 2001; Guha *et al.,* 2001) and Sp1 (Gum *et al.,* 1997; Kovarik *et al.,* 1996; Perrais *et al.,* 2001).

Including C/EBP and Sp1 in the list was additionally reasoned by the fact that these factors are known to be second constituents in the most frequent NF-κB-containing composite elements as they are compiled in the TRANSCompel® database (Wingender *et al.,* 1997). Moreover, these are the types of composite elements known to participate in different kinds of immune response.

4.5.2. Models for LPS and MALP-2

10 TFs were selected as the most important triggers of the early LPS response: AP-1, ETS, Elk-1, NF-κB, ATF2, C/EBP, CREB, NFAT, Sp1 and SRF. The information was taken from TRANSFAC® , TRANSPATH® and www.malp-research.de and supported with additional literature evidence (Kawai *et al.,* 2001 (IRF, NF-κB); Sato *et al.,* 2000 (IRF, NF-κB); Krappmann *et al.,* 2004 (AP-1, NF- κB); Lin *et al.,* 1998 (IRF, NF-κB); Guha and Mackman, 2001 (CREB); Herlaar and Brown, 1999 (ATF-2, CREB); Ling *et al.,* 1998 (SRF and Elk-1); Heidenreich *et al.,* 1999; Dieterich *et al.,* 2003 (SRF)).

The set for the models for LPS-triggered and MALP-2-triggered genes was practically the same, because the triggered pathways are the same except for the MyD-88-independent pathway (Sato *et al.,* 2000). In the model for MyD-88-independent pathway Sp1 was substituted with IRF (Kawai *et al.,* 2001).

4.6. Search for the potential transcription factor binding sites

4.6.1. For promoter model construction

We made this search with the weight matrix approach using the Match™ tool (Kel *et al.,* 2003); the matrices were chosen from the library collected in TRANSFAC® (Kel *et al.,* 2003). For the model construction, the thresholds for the matrix search have been defined individually for each matrix and in such a way that (i) it should yield not less than 80% TP (true positive set, here the set of experimentally proven TFBS from TRANSFAC®); (ii) at least one hit for every searched transcription factor could be found in every sequence of the (+)-training set. The lower border for the thresholds was predefined as 0.80/0.79 (core similarity/ matrix similarity).

4.6.2. In the set of CE-containing sequences (application of distance distribution approach)

The TFBS were searched with the weight matrix approach using the Match™ tool (Kel *et al.,* 2003); the matrices were chosen from the library collected in TRANSFAC® (Matys *et al.,* 2003). The thresholds for the matrix search have been defined individually for each matrix in such a way that every experimentally proven binding site in a CE should be reidentified.

4.7. Identification of pairs

We considered all the coordinates of all potential TF binding sites found by Match™ for each transcription factor (for each set of CE-containing sequences only the TFBS for the constituents of the CE were searched). Further on, we examined all possible combinations of the coordinates, thus revealing all possible pairs in the sequence. Only heterogeneous pairs were considered. The distances were measured between the centers of the sites.

5. SUMMARY

This work describes the development of several new methods to construction of promoter models as one of necessary steps of regulatory networks construction. Deciphering the promoter structure of co-regulated genes enables to obtain information about the pathways of the corresponding signaling. Identification of characteristic features of promoters shows the role of specific transcription factors in triggering the specific response, which in turn sheds light on the signaling pathways activating these transcription factors. Treating reported results of microarray analyses together with other available information about the genes expressed in the cellular systems under consideration, we search for distinguishing features of the promoters of coexpressed genes. The application of such promoter models enables to identify additional candidate genes belonging to the same regulatory network.

Four novel approaches are presented in this work: (i) subtractive approach to matrix generation; (ii) distance distribution approach; (iii) "seed" sets approach; (iv) complementary pairs approach.

These approaches help to solve serious problems in promoter model construction such as the doubtful reliability of positive training sets ("seed" sets approach) and lack of knowledge about the exact signaling pathways triggering the gene expression (complementary pairs approach); the subtractive approach to matrix generation allows to refine positional weight matrices for heterogeneous sets of binding sites, thus to improve the PWM search for single TFBS. Significant improvement of the specificity of promoter analysis has been achieved by applying statistical methods for characterizing TFBS combinations at over-represented distances rather than the mere identification of single potential TFBS (distance distributions approach).

The newly developed methods were applied to the description of four defensive eukaryotic systems in terms of transcription regulation. The obtained models enabled us to gain better insights into the pathways of the corresponding signaling networks.

6. REFERENCES

Akira S and Hemmi H. (2003) Recognition of pathogen-associated molecular patterns by TLR family. Immunol Lett. 85: 85-95

Alkema WB, Johansson O, Lagergren J, Wasserman WW. (2004) MSCAN: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res. 32: W195-W198

Au WC, Moore PA, Lowther W, Juang YT, Pitha PM. (1995) Identification of a member of the interferon regulatory factor family that binds to the interferon-stimulated response element and activates expression of interferon-induced genes. Proc Natl Acad Sci USA. 92: 11657-11661

Bailey TL and Elkan C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, California, 28-36

Bailey TL and Noble WS. (2003) Searching for statistically significant regulatory modules. Bioinformatics. 1: 1-10

Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V. (2002) Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics. 18: 198-199

Becker MN, Diamond G, Verghese MW, Randell SH. (2000) CD14-dependent lipopolysaccharide-induced β-defensin-2 expression in human tracheobronchial epithelium. J Biol Chem. 275: 29731-29736

Ben-Baruch A, Michiel DF, Oppenheim JJ. (1995) Signals and receptors involved in recruitment of inflammatory cells. J Biol Chem. 270: 11703-11706

Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics. 21: 2657-2666.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. (2005) GenBank. Nucleic Acids Res. 33: D34-D38

Berezikov E, Guryev V, Plasterk RH, Cuppen E. (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. Genome Res. 14: 170-178.

Berg OG and von Hippel PH. (1987) Selection of DNA binding sites by regulatory

proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol. 193: 723-750

Bergmann M, Hart L, Lindsay M, Barnes PJ, Newton R. (1998) IkappaBalpha degradation and nuclear factor-kappaB DNA binding are insufficient for interleukin-1beta and tumor necrosis factor-alpha-induced kappaB-dependent transcription Requirement for an additional activation pathway. J Biol Chem. 273: 6607-6610

Berman BP, Nibu Y, Pfeiffer BD, Tomanchak P, Celniker SE, Levine M, Rubin GM, Eisen MB. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci USA. 99: 757-762

Bilke S, Breslin T, Sigvardsson M. (2003) Probabilistic estimation of microarray data reliability and underlying gene expression. BMC Bioinformatics. 4: 40

Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T. (2004) Ensembl 2004. Nucleic Acids Res. 32: D468-D470

Brazma A, Jonassen I, Vilo J, Ukkonen E. (1998) Predicting gene regulatory elements in silico on a genomic scale. Genome Res. 8: 1202-1215

Britigan BE, Railsback MA, Cox CD. (1999) The Pseudomonas aeruginosa secretory product pyocyanin inactivates α1 protease inhibitor: implications for the pathogenesis of cystic fibrosis lung disease. Infect Immun. 67: 1207-1212

Britigan BE, Roeder TL, Rasmussen GT, Shasby DM, McCormick ML, Cox CD. (1992) Interaction of the Pseudomonas aeruginosa secretory products pyocyanin and pyochelin generates hydroxyl radical and causes synergistic damage to endothelial cells. Implications for Pseudomonas-associated tissue injury. J Clin Invest. 90(6): 2187-2196

Buhlmann P and Wyner AJ. (1999) Variable length Markov chains. Ann. Statist. 27: 480-513

Bulyk ML, Johnson PL, Church GM. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. 30: 1255-1261

Cao Z, Xiong J, Takeuchi M, Kurama T, Goeddel DV. (1996) TRAF6 is a signal transducer for interleukin-1. Nature. 383(6599): 443-446

Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics. 21(13): 2933-2942

Chekmenev DS, Haid C, Kel AE. (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. Nucleic Acids Res. 33: W432-W437

Chen M, Lin S, Hofestaedt R. (2004) STCDB: Signal Transduction Classification Database. Nucleic Acids Res. 32: D456-D458

Cheng G, Nazar AS, Shin HS, Vanguri P, Shin ML. (1998) IP-10 gene transcription by virus in astrocytes requires cooperation of ISRE with adjacent B site but not IRF-1 or viral transcription. J. Interferon Cytokine Res. 11: 987-997

Chiang DY, Moses AM, Kellis M, Lander ES, Eisen MB. (2003) Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. Genome Biol. 4: R43

Cobb LM, Mychaleckyj JC, Wozniak DJ, Lopez-Boado YS. (2004) Pseudomonas aeruginosa flagellin and alginate elicit very distinct gene expression patterns in airway epithelial cells: implications for cystic fibrosis disease. J Immunol. 173(9): 5659-5670

Corcoran DL, Feingold E, Dominick J, Wright M, Harnaha J, Trucco M, Giannoukakis N, Benos PV. (2005) Footer: a quantitative comparative genomics method for efficient recognition of cis-regulatory elements. Genome Res. 15(6): 840-847

Costerton, JW, Stewart, PS, Greenberg, EP. (1999) Bacterial biofilms: a common cause of persistent infections. Science. 284: 1318-1322

Crowley EM, Roeder K, Bina M.(1997) A statistical model for locating regulatory regions in genomic DNA. J.Mol.Biol. 268: 8-14

Davuluri RV, Grosse I, Zhang MQ. (2001) Computational identification of promoters and first exons in the human genome. Nat Genet. 29(4): 412-417

Deiters U, Muhlradt PF. (1999) Mycoplasmal lipopeptide MALP-2 induces the chemoattractant proteins macrophage inflammatory protein 1alpha (MIP-1alpha), monocyte chemoattractant protein 1, and MIP-2 and promotes leukocyte infiltration in mice. Infect Immun. 67(7) : 3390-3398

Diamond G, Jones DE, Bevins CL. (1993) Airway epithelial cells are the site of expression of a mammalian antimicrobial peptide gene. Proc Natl Acad Sci USA. 90:

4596-4600

Diamond G, Kaiser V, Rhodes J, Russell JP, Bevins C. (2000) Transcriptional regulation of b-defensin gene expression in tracheal epithelial cells. Infection and immunity. 68: 113-119

Dieterich C, Herwig R, Vingron M. (2003) Exploring potential target genes of signaling pathwas by predicting conserved transcription factor binding sites. Bioinformatics. 19: suppl2, ii50-ii56

DiMango E, Ratner AJ, Bryan R, Tabibi S,, Prince A. (1998) Activation of NF-κB by adherent Pseudomonas aeruginosa in normal and cystic fibrosis respuratory epithelial cells. J Clin Invest. 101: 2598-2606

Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, Tainsky MA. (2003) Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics. 19: 348-359

Duret L, Bucher P. (1997) Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol. 7: 399-406

Eskin E, Pevzner PA. (2002) Finding composite regulatory patterns in DNA sequences. Bioinformatics.18: Suppl 1, S354-S363

Etzold T, Ulyanov A, Argos P. (1996) SRS: information retrieval system for molecular biology data banks. Methods Enzymol. 266; 114-128

Fickett JW, Wasserman WW. (2000) Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol. 11: 19-24

Fickett JW. (1996) Finding genes by computer: the state of the art. Trends Genet. 8: 316-320

Fitzgerald KA, Rowe DC, Barnes BJ, Caffrey DR, Visintin A, Latz E, Monks B, Pitha PM, Golenbock DT. (2003) LPS-TLR4 signaling to IRF-3/7 and NF-kappaB involves the toll adapters TRAM and TRIF. J Exp Med. 198(7): 1043-1055. Erratum in: J Exp Med. (2003) 198(9): following 1450

Frech K, Danescu-Mayer J, Werner T. (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J Mol Biol. 270: 674-687

Frech K, Quandt K, Werner T. (1997) Finding protein-binding sites in DNA sequences: the next generation.Trends Biochem Sci. 3:103-104

Frith MC, Hansen U, Weng Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics. 10: 878-889

Frith MC, Spouge JL, Hansen U, Weng Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. Nucleic Acids Res 30(14): 3214-3224

Fritz G and Kaina B. (2001) Transcriptional activation of the small GTPase rhoB by genotoxic stress is regulated via a CCAAT element. Nucleic Acids Res. 29: 792-798

Ghosh D. (1998) OOTFD (Object-Oriented Transcription Factors Database): an object-oriented successor to TFD. Nucleic Acids Res. 26:360-362

Gnad R, Kaina B, Fritz G. (2001) Rho GTPases are involved in the regulation of NF-kB by genotoxic stress. Exp Cell Res 264: 244-249

Goodrich JA, Schwartz ML, McClure WR. (1990) Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for Escherichia coli integration host factor (IHF). Nucleic Acids Res. 18: 4993-5000

Grabe N. (2002) AliBaba2: context specific identification of transcription factor binding sites. In Silico Biol. 2(1): S1-15

Guha M., Mackman N. (2001) LPS induction of gene expression in human monocytes. Cell. Signal. 13: 85-94

Guha M, O'Connell M A, Pawlinski R, Hollis A, McGovern P, Yan S F, Stern D, Mackman N. (2001) Lipopolysaccharide activation of the MEK-ERK1/2 pathway in human monocytic cells mediates tissue factor and tumor necrosis factor alpha expression by inducing Elk-1 phosphorylation and Egr-1 expression. Blood, 98: 1429-1439

GuhaThakurta D and Stormo GD. (2001) Identifying target sites for cooperatively binding factors.Bioinformatics. 7: 608-621

Gum JR Jr, Hicks JW, Kim YS. (1997) Identification and characterization of the MUC2 (human intestinal mucin) gene 5'-flanking region: promoter activity in cultured cells. Biochem J 325: 259-267

Hannenhalli S, Levy S. (2002) Predicting transcription factor synergism. Nucleic Acids Res 30: 4278-4284

Harder J, Meyer-Hoffert U, Teran L M, Schwichtenberg L, Basrtels J, Maune S, Schroeder J-M. (2000) Mucoid Pseudomonas aeruginosa, TNFα, and IL-1β, but not IL-6, induce human β-defensin-2 in respiratory epithelia. Am J Respir Cell Mol Biol 22: 714-721

Hardison RC. (2003) Comparative Genomics. PLoS Biol. 1: E58

Harr R, Haggstrom M, Gustafsson P. (1983) Search algorithm for pattern match analysis of nucleic acid sequences. Nucleic Acids Res. 11: 2943-2957

Heidenreich O., Neininger A., Schratt G., Zinck R., Cahill M.A., Engel K., Kotlyarov A., Kraft R., Kostka S., Gaestel M., Nordheim A. (1999) MAPKAP kinase 2 phosphorylates serum response factor in vitro and in vivo. J Biol Chem.274: 14434-14443

Herlaar E. and Brown Z. (1999) p38 MAPK signalling cascades in inflammatory disease. Mol. Med. Today. 5: 439-447

Hertz GZ, Hartzell GW 3rd, Stormo GD. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput Appl Biosci. 6(2): 81-92

Hertz GZ and Stormo GD. (1995) Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In: Proceedings of the Third International Conference on Bioinformatics and Genome Research (H.A. Lim, and C.R. Cantor, editors). World Scientific Publishing Co., Ltd. Singapore, 201-216

Higo K, Ugawa Y, Iwamoto M, Korenaga T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 27: 297-300

Hoffmann JA, Kafatos FC, Janeway CA, Ezekowitz RA. (1999) Phylogenetic perspectives in innate immunity. Science. 284(5418):1313-1318

Hoiby N. (2002) New antimicrobials in the management of cystic fibrosis. J Antimicrob Chemother. 2: 235-238

Hughes JD, Estep PW, Tavazoie S, Church GM. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol. 296(5): 1205-1214

Ichikawa JK, Norris A, Bandera MG, Geiss GK, van ‘t Wout AB, Bumgarner R, Lory S. (2000) Interaction of Pseudomonas aeruginosa with epithelial cells: Identification of differentially regulated genes by expression microarray analysis of human cDNAs. Proc Natl Acad Sci USA, 97: 9659-9664

Imundo L, Barasch J, Prince A, Al-Awqati Q. (1995) Cystic fibrosis epithelial cells have a receptor for pathogenic bacteria on their apical surface. Proc Natl Acad Sci USA. 92(7): 3019-3023. Erratum in: Proc Natl Acad Sci USA (1995) 92(24): 11322

Janeway CA Jr and Medzhitov R. (2002) Innate immune recognition. Annu Rev Immunol. 20:197-216

Jefferies CA and O'Neill LA. (2004) Bruton's tyrosine kinase (Btk)-the critical tyrosine kinase in LPS signalling? Immunol Lett. 92(1-2): 15-22

Jegga AG, Gupta A, Gowrisankar S, Deshmukh MA, Connolly S, Finley K, Aronow BJ. (2005) CisMols Analyzer: identification of compositionally similar cis-element clusters in ortholog conserved regions of coordinately expressed genes. Nucleic Acids Res. 33: W408-W411. Erratum in: Nucleic Acids Res. 2005.33: 4377

Kim J, Seo J, Lee YS and Kim S. (2005) TFExplorer: Integrated Analysis Database for predicted Transcription Regulatorory Elements. Bioinformatics, 21: 548-550

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 33: D428-D432

Kankainen M and Holm L. (2004) POBO, transcription factor binding site verification with bootstrapping. Nucleic Acids Res. 32: W222-W229

Kankainen M and Holm L. (2005) POCO: discovery of regulatory patterns from promoters of oppositely expressed gene sets. Nucleic Acids Res. 33: W427-W431

Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Diez FG, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R. (2005) The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 33: D29-D33

Kaufmann A, Muhlradt PF, Gemsa D, Sprenger H. (1999) Induction of cytokines and chemokines in human monocytes by Mycoplasma fermentans-derived lipoprotein MALP-2. Infect Immun. 67(12): 6303-6308

Kawai T, Takeuchi O, Fujita T, Inoue J, Muhlradt PF, Sato S, Hoshino K, Akira S. (2001) Lipopolysaccharide stimulates the MyD88-independent pathway and results in activation of IFN-regulatory factor 3 and the expression of a subset of lipopolysaccharide-inducible genes. J Immunol. 167(10): 5887-5894

Kel A, Kel-Margoulis O, Babenko V, Wingender E. (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. J Mol Biol. 288(3): 353-376

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 31: 3576-3579

Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ. (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. J Mol Biol. 309: 99-120

Kel AE, Kondrakhin YV, Kolpakov Ph.A., Kel OV, Romashenko AG, Wingender E, Milanesi L and Kolchanov NA. (1995) Computer tool FUNSITE for analysis of eucaryotic regulatory genomic sequences. Intell.Syst.Mol.Biol. 3: 197-205

Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E. (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. Nucleic Acids Res. 30: 332-334

Kel-Margoulis OV, Tchekmenev D, Kel AE, Goessling E, Hornischer K, Lewicki-Potapov B, Wingender E. (2003) Composition-sensitive analysis of the human genome for regulatory signals In Silico Biol. 3: 0017

Klingenhoff A, Frech K, Werner T. (2002) Regulatory modules shared within gene classes as well as across gene classes can be detected by the same in silico approach. In Silico Biol. 2: S17-S26

Ko YH, Delannoy M, Pedersen PL. (1997) Cystic fibrosis, lung infections, and a human tracheal antimicrobial peptide (hTAP). FEBS letters 405: 200-208

Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res. 30: 312-317

Kondrakhin YV, Kel AE, Kolchanov NA, Romashchenko AG, Milanesi L. (1995) Eukaryotic promoter recognition by binding sites for transcription factors. Comput Appl Biosci. 11(5): 477-488

Kovarik A, Lu PJ, Peat N, Morris J, Taylor-Papadimitriou J. (1996) Two GC boxes (Sp1 sites) are involved in regulation of the activity of the epithelium-specific MUC1 promoter. J Biol Chem.271: 8140-8147

Kramer-Hammerle S, Hahn A, Brack-Werner R, Werner T. (2005) Elucidating effects of long-term expression of HIV-1 Nef on astrocytes by microarray, promoter, and literature analyses. Gene. 358: 31-38

Krappmann D, Wegener E, Sunami Y, Esen M, Thiel A, Mordmuller B, Scheidereit C. (2004) The IkappaB kinase complex and NF-kappaB act as master regulators of lipopolysaccharide-induced gene expression and control subordinate activation of AP-1. Mol Cell Biol. 24(14): 6488-6500

Krivan HC, Roberts DD, Ginsburg V. (1988) Many pulmonary pathogenic bacteria bind specifically to the carbohydrate sequence GalNAc beta 1-4Gal found in some glycolipids. Proc Natl Acad Sci USA. 85(16): 6157-6161

Krivan W, Wasserman WW. (2001) A predictive model for regulatory sequences directing liver-specific transcription. Genome Res 11: 1559-1566

Krull M, Voss N, Choi C, Pistor S, Potapov A, Wingender E. (2003) TRANSPATH: an integrated database on signal transduction and a tool for array analysis. Nucleic Acids Res. 31: 97-100

Kus JV, Tullis E, Cvitkovitch DG, Burrows LL. (2004) Significant differences in type IV pilin allele distribution among Pseudomonas aeruginosa isolates from cystic fibrosis (CF) versus non-CF patients.Microbiology. 150: 1315-1326

Lau GW, Hassett DJ, Ran H, Kong F. (2004) The role of pyocyanin in Pseudomonas aeruginosa infection.Trends Mol Med. 12: 599-606

Lau GW, Ran H, Kong F, Hassett DJ, Mavrodi D. (2004) Pseudomonas aeruginosa pyocyanin is critical for lung infection in mice.Infect Immun. 72:4275-4278

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 262(5131): 208-214

Lee ML, Kuo FC, Whitmore GA, Sklar J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA. 97: 9834-9839

Lee CG, Jenkins NA Gilbert DJ Copeland NG O'Brien WE. (1995) Cloning and analysis of gene regulation of a novel LPS-inducible cDNA. Immunogenetics. 41: 263-270

Leidal KG, Munson KL, Denning GM. (2001) Small molecular weight secretory factors from Pseudomonas aeruginosa have opposite effects on IL-8 and RANTES expression by human airway epithelial cells Am J Respir Cell Mol Biol. 25: 186-195

Lemer C, Antezana E, Couche F, Fays F, Santolaria X, Janky R, Deville Y, Richelle J, Wodak SJ. (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. Nucleic Acids Res. 32: D443-D448

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze P, Rombauts S. (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 30: 325-327

Levy S, Hannenhalli S, Workman C. (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics. 17: 871-877

Li J-D, Feng W, Gallup M, Kim J-H, Kim J, Kim Y, Basbaum C. (1998) Activation of NF-kB via a Src-dependent Ras-MAPKpp90rsk pathway is required for Pseudomonas aeruginosa-induced micin overproductionin epithelial cells. Proc Natl Acad Sci USA, 95: 5718-5723

Lin R, Heylbroeck C, Pitha PM, Hiscott J. (1998) Virus-dependent phosphorylation of the IRF-3 transcription factor regulates nuclear translocation, transactivation potential, and proteasome-mediated degradation.Mol Cell Biol. 18: 2986-2996

Lin R, Heylbroeck C, Genin P, Pitha PM, Hiscott J. (1999). Essential role of interferon regulatory factor-3 in direct activation of RANTES chemokine transcription. Mol. Cell. Biol. 19: 959-966

Ling Y, West AG, Roberts EC, Lakey JH, Sharrocks AD. (1998) Interaction of transcription factors with serum response factor. Identification of the Elk-1 binding surface J Biol Chem. 273: 10506-10514

Liu L, Wang L, Jia HP, Zhao C, Heng HH, Schutte BC, McCray PB Jr, Ganz T. (1998) Structure and mapping of the human beta-defensin HBD-2 gene and its expression at sites of inflammation. Gene. 222: 237-244

Liu X, Brutlag DL, Liu JS. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput.: 127-138

Liu XS, Brutlag DL, Liu JS. (2002) An algorithm for finding protein DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol. 20(8): 835-839

Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S. (2004) Conservation of eukaryotic regulatory elements and their identification using comparative genomics. Genome Res. 14: 451-458

Lukashin AV, Anshelevich VV, Amirikyan BR, Gragerov AI, Frank-Kamenetskii MD. (1989) Neural network models for promoter recognition. J Biomol Struct Dyn. 6: 1123-1133

Lyon GD, Newton AC, Marshall B. (2002) The need for a standard nomenclature for gene classification and a generic, automated tool to assist in hypothesis formulation in cell signalling. Molecular Plant Pathology. 3 (2): 103-109

Mah TC, O'Toole GA (2001) Mechanisms of biofilm resistance to antimicrobial agents. Trends Microbiol. 9: 34-39

Mai GT, Seow WK, Pier GB, McCormack JG, Thong YH. (1993) Suppression of lymphocyte and neutrophil functions by Pseudomonas aeruginosa mucoid

exopolysaccharide (alginate): reversal by physicochemical, alginase, and specific monoclonal antibody treatments. Infect Immun. 61: 559-564

Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA. (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucleic Acids Res. 31: 6016-6026

Markstein M, Markstein P, Markstein V, Levine MS. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci USA, 99: 763-768

Mathee K, Ciofu O, Sternberg C, Lindum PW, Campbell JI, Jensen P, Johnsen AH, Givskov M, Ohman DE, Molin S, Hoiby N, Kharazmi A. (1999) Mucoid conversion of Pseudomonas aeruginosa by hydrogen peroxide: a mechanism for virulence activation in the cystic fibrosis lung. Microbiology. 145: 1349-1357

Matis S, Xu Y, Shah M, Guan X, Einstein JR, Mural R, Uberbacher E. (1996) Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. Comput Chem. 20(1): 135-140

Mattick, JS (2002) Type IV pili and twitching motility. Annu Rev Microbiol. 56: 89-314

Mattick JS, Whitchurch CB, Alm RA (1996) The molecular genetics of type-4 fimbriae in Pseudomonas aeruginosa: a review. Gene.179: 147-155

Matys V, Fricke E, Geffers R, Gößling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31: 374-378

McNamara N, Khong A, McKemy D, Caterina M, Boyer J, Julius D, Basbaum, C. (2001) ATP transduces signals from ASGM1, a glycolipid that functions as a bacterial receptor. Proc Natl Acad Sci USA. 98: 9086-9091

Medzhitov R and Janeway CA Jr. (1997) Innate immunity: impact on the adaptive immune response. Curr Opin Immunol. 1: 4-9

Merienne K, Pannetier S, Harel-Bellan A, Sassone-Corsi P. (2001) Mitogen-regulated RSK2-CBP interaction controls their kinase and acetylase activities. Mol Cell Biol. 20: 7089-7096

Messeguer X, Escudero R, Farre D, Nunez O, Martinez J, Alba MM. (2002) PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics. 2: 333-334

Mori N, Oishi K, Sar B, Mukaida N, Nagatake T, Matsushima K, Yamamoto N. (1999) Essential role of transcription factor nuclear factor-kappaB in regulation of interleukin-8 gene expression by nitrite reductase from Pseudomonas aeruginosa in respiratory epithelial cells. Infect Immun. 67: 3872-3878

Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. (2003) Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol Biol. 3: 19

Mühlradt PF, Tsai H, Conradt P. (1986) Effects of pyocyanine, a blue pigment from Pseudomonas aeruginosa, on separate steps of T cell activation: interleukin 2 (IL 2) production, IL 2 receptor formation, proliferation and induction of cytolytic activity. Eur J Immunol. 4: 434-440

Müller PK, Krohn K, Mühlradt PF. (1989) Effects of pyocyanine, a phenazine dye from Pseudomonas aeruginosa, on oxidative burst and bacterial killing in human neutrophils. Infect Immun. 57: 2591-2596

Münch R, Hiller K, Barg H, Heldt D, Linz S, Wingender E, Jahn D. (2003) PRODORIC: prokaryotic database of gene regulation. Nucleic Acids Res. 31: 266-269

Navarro L and David M. (1999) p38-dependent activation of interferon regulatory factor 3 by lipopolysaccharide. J Biol Chem. 274(50): 35535-35538

Navarro L, Mowen K, Rodems S, Weaver B, Reich N, Spector D, David M. (1998) Cytomegalovirus activates interferon immediate-early response gene expression and an interferon regulatory factor 3-containing interferon-stimulated response element-binding complex. Mol Cell Biol. 7:3796-3802

Ohler U, Harbeck S, Niemann H, Noth E, Reese MG. (1999) Interpolated markov chains for eukaryotic promoter recognition. Bioinformatics. 5: 362-369

Ohler U and Niemann H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches.Trends Genet. 2: 56-60

Ohmori Y and Hamilton TA. (1993) Cooperative interaction between interferon (IFN) stimulus response element and B sequence motifs controls IFN-- and lipopolysaccharide-stimulated transcription from the murine IP-10 promoter. J Biol Chem. 268: 6677

O'Malley YQ, Abdalla MY, McCormick ML, Reszka KJ, Denning GM, Britigan BE. (2003a) Subcellular localization of Pseudomonas pyocyanin cytotoxicity in human lung epithelial cells. Am J Physiol Lung Cell Mol Physiol. 284: L420-L430.

O'Malley YQ, Reszka KJ, Rasmussen GT, Abdalla MY, Denning GM, Britigan BE.

(2003b) The Pseudomonas secretory product pyocyanin inhibits catalase activity in human lung epithelial cells. Am J Physiol Lung Cell Mol Physiol. 285: L1077-L1086

O'Neill MC. (1991) Training back-propagation neural networks to define and detect DNA-binding sites. Nucleic Acids Res. 19: 313-318

Orlov YL and Potapov VN. (2000) Determining Markov model of genetical texts by stochastic complexity estimation. BGRS, Novosibirsk, 71-77

Orlov YL, Filippov VP, Potapov VN, Kolchanov NA. (2002) Construction of stochastic context trees for genetic texts. In Silico Biol. 2(3): 233-247

Oshiumi H, Matsumoto M, Funami K, Akazawa T, Seya T. (2003) TICAM-1, an adaptor molecule that participates in Toll-like receptor 3-mediated interferon-beta induction.Nat Immunol. 2:161-167

Palsson-McDermott EM and O'Neill LA. (2004) Signal transduction by the lipopolysaccharide receptor, Toll-like receptor-4. Immunology. 113(2): 153-162

Pan WA. (2002) Comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics. 18: 546-554

Passador L, Cook JM, Gambello MJ, Rust L, Iglewski BH. (1993) Expression of Pseudomonas aeruginosa virulence genes requires cell-to-cell communication. Science. 260(5111): 1127-1130

Pennacchio LA, Rubin EM. (2003) Comparative genomic tools and databases: providing insights into the human genome. J Clin Invest. 111: 1099-1106

Perier RC, Junier T, Bonnard C, Bucher P. (1999) The Eukaryotic Promoter Database (EPD): recent developments. Nucleic Acids Res. 27:307-309

Perrais M, Pigny P, Ducourouble MP, Petitprez D, Porchet N, Aubert JP, Van Seuningen I. (2001) Characterization of human mucin gene MUC4 promoter: importance of growth factors and proinflammatory cytokines for its regulation in pancreatic cancer cells. J Biol Chem. 76: 30923-30933

Pickert L, Reuter I, Klawonn F, Wingender E. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering.Bioinformatics..14(3): 244-251

Poluliakh N, Takagi T, Nakai K. Melina. (2003) motif extraction from promoter regions of potentially co-regulated genes. Bioinformatics. 19(3): 423-424

Prestridge D. (1995) Predicting PolII promoter sequences using transcription factor binding sites. J Mol Biol. 249: 923-932

Prince A. (1992) Adhesins and receptors of Pseudomonas aeruginosa associated with infection of the respiratory tract. Microb Pathog. 4: 251-260

Pritchard CC, Hsu L, Delrow J, Nelson PS. (2001) Project normal: defining normal variance in mouse gene expression. Proc Natl Acad Sci USA. 98: 13266-13271

Pruitt KD, Tatusova T, Maglott DR. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33: D501-504

Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. Nat Biotechnol. 4: 435-439

Quandt K, Frech K, Karas H, Wingender E, Werner T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23: 4878-4884

Ran H, Hassett DJ, Lau GW. (2003) Human targets of Pseudomonas aeruginosa pyocyanin. Proc Natl Acad Sci USA. Nov 25.100(24):14315-20. Epub (2003) Nov 6

Ratner A, Bryan R, Weber A, Nguyen S, Barnes D, Pitt A, Gelber S, Cheung A,, Prince A. (2001) Cystic fibrosis pathogens activate Ca2+-dependent mitogen-activated protein kinase signaling pathways in airway epithelial cells. J Biol Chem. 276: 19267-19275

Reich N, Evans B, Levy D, Fahey D, Knight E Jr, Darnell JE Jr. (1987) IFN-induced transcription of a gene encoding a 15-kDa protein depends on an upstream enhancer element. Proc. Natl. Acad. Sci. USA. 84: 6394-6398

Rietschel ET, Kirikae T, Schade FU, Mamat U, Schmidt G, Loppnow H, Ulmer AJ, Zahringer U, Seydel U, Di Padova F, Schreier M and Brade H. (1994) Bacterial endotoxin: molecular relationships of structure to activity and function. FASEB J. 2: 217-225

Rissanen J. (1983) A universal data compression system. IEEE Trans. Inform. Theory. 29: 656-664

Sacht G, Marten A, Deiters U, Sussmuth R, Jung G, Wingender E, Muhlradt PF. (1998) Activation of nuclear factor-kappaB in macrophages by mycoplasmal lipopeptides. Eur J Immunol. 12: 4207-4212

Salgado H, Santos A, Garza-Ramos U, van Helden J, Diaz E, Collado-Vides J. (1999) RegulonDB (version 2.0): a database on transcriptional regulation in Escherichia coli. Nucleic Acids Res. 27: 59-60

Salzberg SL, Delcher AL, Kasif S, White O. (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26: 544-548

Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H. (1999) Interpolated Markov models for eukaryotic gene finding. Genomics. 59: 24-31

Sandelin A, Wasserman WW, Lenhard B. (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res. 32: W249-W252

Sar B, Oishi K, Matsushima K, Nagatake T. (1999) Induction of interleukin 8 (IL-8) production by Pseudomonas nitrite reductase in human alveolar macrophages and epithelial cells. Microbiol Immunol. 43: 409-417

Sar B, Oishi K, Wada A, Hirayama T, Matsushima K, Nagatake T. (2000) Induction of monocyte chemoattractant protein-1 (MCP-1) production by Pseudomonas nitrite reductase in human pulmonary type II epithelial-like cells. Microb Pathog. 28: 17-23

Sar B, Oishi K, Wada A, Hirayama T, Matsushima K, Nagatake T. (1999) Nitrite reductase from Pseudomonas aeruginosa released by antimicrobial agents and complement induces interleukin-8 production in bronchial epithelial cells. Antimicrob Agents Chemother. 43: 794-801

Sato S, Nomura F, Kawai T, Takeuchi O, Muhlradt PF, Takeda K, Akira S. (2000) Synergy and cross-tolerance between toll-like receptor (TLR) 2- and TLR4-mediated signaling pathways. J Immunol. 165: 7096-7101

Schacherer F, Choi C, Götze U, Krull M, Pistor S, Wingender E. (2001) The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. Bioinformatics. 11: 1053-1057

Scherf M, Klingenhoff A, Werner T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol. 297: 599-606

Schroeder TH, Zaidi T, Pier GB. (2001) Lack of adherence of clinical isolates of Pseudomonas aeruginosa to asialo-GM(1) on epithelial cells. Infect Immun. 69: 719-729

Seifert M, Scherf M, Epple A, Werner T. (2005) Multievidence microarray mining. Trends Genet. 10: 553-8

Shelest E, Kel AE, Goessling E, Wingender E. (2003) Prediction of potential C/EBP/NF-kappaB composite elements using matrix-based search methods. In Silico Biol. 3(1-2): 71-79.

Shelest E and Wingender E. (2005) Construction of predictive promoter models on the example of antibacterial response of human epithelial cells. Theor. Biol. Med. Model.

2(1): 2

Sheth HB, Lee KK, Wong WY, Srivastava G, Hindsgaul O, Hodges RS, Paranchych W, Irvin RT. (1994) The pili of Pseudomonas aeruginosa strains PAK and PAO bind specifically to the carbohydrate sequence beta GalNAc(1-4)beta Gal found in glycosphingolipids asialo-GM1 and asialo-GM2. Mol Microbiol. Feb.11(4):715-723

Shi W and Sun H. (2002) Type IV pilus-dependent motility and its possible role in bacterial pathogenesis. Infect Immun. 70: 1-4

Shuman JD, Cheong J, Coligan JE. (1997) ATF-2 and C/EBPalpha can form a heterodimeric DNA binding complex in vitro. Functional implicationsfor transcriptional regulation. J Biol Chem. 272(19): 12793-12800

Silverman N, Maniatis T. (2001) NF-kappaB signaling pathways in mammalian and insect innate immunity.Genes Dev. 15(18): 2321-2342

Singh PK, Jia HP, Wiles K, Hesselberth J, Liu L, Conway BA, Greenberg EP, Valore EV, Welsh MJ, Ganz T, Tack BF, McCray PB Jr. (1998) Production of beta-defensins by human airway epithelia. Proc Natl Acad Sci USA. 95: 14961-14966

Sivakumaran S, Hariharaputran S, Mishra J, Bhalla US. (2003) The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks. Bioinformatics. 19: 408-415

Smith RS, Fedyk ER, Springer TA, Mukaida N, Iglewski BH, Phipps RP. (2001) IL-8 production in human lung fibroblasts and epithelial cells activated by the Pseudomonas aeruginosa autoinducer N-3-oxodododecanoyl homoserine lactone is transcriptionally regulated by NF-κB and activator protein-2. J immunol. 167: 366-374

Smith JB and Herschman HR. (1996) The glucocorticoid-attenuated response genes GARG-16, GARG-39, and GARG-49/IRG2 encode inducible proteins containing multiple tetratricopeptide repeat domains. Arch. Biochem. Biophys. 330: 290-300

Staden R. (1984) Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 12: 505-519

Stevenson MA, Zhao MJ, Asea A, Coleman CN, Calderwood SK. (1999) Salicylic acid and aspirin inhibit the activity of RSK2 kinase and repress RSK2-dependent transcription of cyclic AMP response element binding protein- and NF-kappa B-responsive genes. J Immunol. 163(10): 5608-5616

Stormo GD and Hartzell GW 3rd. (1989) Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci USA. 86: 1183-1187

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. 10: 2997-3011

Stormo GD. (1990) Consensus patterns in DNA. Methods Enzymol..183: 211-221.

Stormo GD. (2000) DNA binding sites: representation and discovery. Bioinformatics. 1: 16-23

Suzuki Y, Yamashita R, Nakai K, Sugano S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res. 30: 328-331

Suzuki Y, Yamashita R, Sugano S, Nakai K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Res. 32: D78-D81.

Takai-Igarashi T and Kaminuma T. (1999) A pathway finding system for the cell signaling networks database. In Silico Biol..1(3): 129-146

Takai-Igarashi T and Mizoguchi R. (2004) Cell signaling networks ontology. In Silico Biol. 2004. 4(1): 81-87

Takai-Igarashi T, Nadaoka Y, Kaminuma T. (1998) A database for cell signaling networks. J Comput Biol. 4: 747-754

Takeda K and Akira S. (2004) TLR signaling pathways. Semin Immunol. 1: 3-9

Takeuchi O and Akira S. (2001) Toll-like receptors. their physiological role and signal transduction system.Int Immunopharmacol. 4: 625-635

Tasheva ES, Klocke B, Conrad GW. (2004) Analysis of transcriptional regulation of the small leucine rich proteoglycans. Mol Vis. 10:758-772

Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. (2002) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res. 30: 27-30

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. (1999) Systematic determination of genetic network architecture. Nat Genet. 3: 281-285

Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics. 12: 1113-1122

Thompson W., E. C. Rouchka and C. E. Lawrence (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res. 31: 3580-3585

Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE. (2004) Decoding Human Regulatory Circuits. Genome Research 14: 1967-1974

Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M. (1997) Analysis of the

distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. J Mol Biol. 266: 231-245

Tsunoda T and Takagi T. (1999) Estimating transcription factor bindability on DNA. Bioinformatics. 7-8: 622-630

Vadigepalli R, Chakravarthula P, Zak DE, Schwaber JS, Gonye GE. (2003) PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification. OMICS. 7: 235-252

Vaidyanathan H and Ramos JW. (2003) RSK2 activity is regulated by its interaction with PEA-15. J Biol Chem. 278(34): 32367-32372

van Helden J, Andre B, Collado-Vides J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol. 281(5): 827-842

van Helden J. (2003) Regulatory sequence analysis tools. Nucleic Acids Res. 31: 3593-3596

Voynow JA, Young LR, Wang Y, Horger T, Rose MC, Fischer BM. (1999) Neutrophil elestase increases MUC5AC mRNA and protein expression in respiratory epithelial cells. Am J Physiol. 276: L835-L843

Wagner A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics. 7-8: 776-784

Walsh DE, Greene CM, Carroll TP, Taggard CC, Gallagher PM, O'Neill SJ, McElvaney NG. (2001) Interleukin-8 up-regulation by neutrophil elastase is mediated by MyD88/IRAK/TRAF-6 in human bronchial epithelium. J Biol Chem. 276: 35494-35499

Wang G, Yu T, Zhang W. (2005) WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. Nucleic Acids Res. 33: W412-W416

Wasserman WW and Fickett JW. (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol. 278: 167-181

Wathelet MG, Clauss IM, Nols CB, Content J, Huez GA. (1987) New inducers revealed by the promoter sequence analysis of two interferon-activated human genes. Eur. J. Biochem. 169: 313-321

Wathelet MG, Clauss IM, Nols CB, Content J, Huez GA. (1988) Regulation of two interferon-inducible human genes by interferon, poly(rI), poly(rC), and viruses. Eur. J. Biochem. 174: 323-329

Werner T, Fessele S, Maier H, Nelson PJ. (2003) Computer modeling of promoter organization as a tool to study transcriptional coregulation. FASEB J. 17: 1228-1237

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 33: D39-D45

Whitsett JA, Bachurski CJ, Barnes KC, Bunn PA, Jr., Case LM, Cook DN, Crooks D, Duncan MW, Dwyer-Nield L, Elston RC, Fessler MB, Franklin WA, Friedman N, Garcia JGN, Geraci MW, Glasgow C, Glasser SW, Hardie WD, Henning LM, Johnson GL, Kawkitinarong K, Keith RL, Korfhagen TR, Leikauf GD, Liggett SB, Malcolm KC, Malkinson AM, Mariani TJ, McDowell SA, McGraw DW, Medvedovic M, Moss J, Nogee LM, Nonas S, Pacheco-Rodriguez G, Palmer LJ, Peters DG, Prows DR, Redline S, Regev A, Sartor MA, Schwartz DA, Silverman EK, Steagall WK, Stearman RS, Taveira-DaSilva A, Tichelaar JW, Tomlinson CR, Tsukada K, Weaver TE, Wert SE, Wesselkamper SC, Worthen GS, Xu Y, Zerbe L, Zhang Y, Zhang Y, Choi AMK, and Kaminski N. (2004) Functional Genomics of Lung Disease.Am. J. Respir. Cell Mol. Biol. 31: S1-S81

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüß M, Reuter I, Schacherer F. (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. 28: 316-319

Wingender E, Kel AE, Kel OV, Karas H, Heinemeyer T, Dietze P, Knueppel R, Romaschenko AG, Kolchanov NA. (1997) TRANSFAC, TRRD and COMPEL: towards a frederated database system on transcriptional regulation. Nucleic Acids Res 25: 265-268

Wingender, E. (1997) Classification scheme of eukaryotic transcription factors. Molekularnaya Biologiya 31, 584-600. Mol. Biol. Engl. Tr. (1997) 31: 483-497

Workman C and Stormo GD. (2000) ANN-Spec: A method for discovering transcription factor binding sites with improved specificity. Proc. Pacific Symposium on Biocomputing 2000

Wozniak DJ and Keyser R. (2004) Effects of subinhibitory concentrations of macrolide antibiotics on Pseudomonas aeruginosa. Chest. 125:62S-69S. quiz 69S

Yamamoto M, Sato S, Hemmi H, Uematsu S, Hoshino K, Kaisho T, Takeuchi O, Takeda

K, Akira S. (2003) TRAM is specifically involved in the Toll-like receptor 4-mediated MyD88-independent signaling pathway. Nat Immunol. 11: 1144-1150

Yamamoto M, Sato S, Mori K, Hoshino K, Takeuchi O, Takeda K, Akira S. (2002) Cutting edge: a novel Toll/IL-1 receptor domain-containing adapter that preferentially activates the IFN-beta promoter in the Toll-like receptor signaling. J Immunol. 169: 6668-6672

Yamamoto M, Takeda K, Akira S. (2004) TIR domain-containing adaptors define the specificity of TLR signaling. Mol Immunol. 40: 861-868

Yellboina S, Seshadri J, Kumar MS, Ranjan A. (2004) PredictRegulon: A web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. Nucleic Acids Res. 32: W318-W320

Zhang G and Ghosh S. (2001) Toll-like receptor–mediated NF-kB activation: a phylogenetically conserved paradigm in innate immunity. J Clin Invest. 107: 13-19

Zhu J and Zhang MQ. (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics.7-8: 607-611

APPENDIX 1. Subtractive approach.

A. C/EBP matrices obtained with the subtractive approach to matrix generation

2002:

Name: CEBP_comp
Binding Matrix:

A	C	G	T	
0	5	1	2	C
0	4	3	1	B
3	3	0	2	H
3	3	0	2	H
4	3	1	0	V
2	1	3	2	N
3	1	4	0	V
0	0	0	8	**T**
0	0	0	8	**T**
1	0	6	1	**G**
0	8	0	0	**C**
7	0	0	1	**A**
2	5	0	1	**C**
6	1	0	1	**A**
6	0	1	1	**A**
2	3	2	1	**N**

Name: CEBP_alternative
Binding Matrix:

A	C	G	T	
0	14	0	5	C
10	5	3	1	A
0	0	0	19	T
0	0	0	19	T
3	0	6	10	K
0	19	0	0	C
0	11	8	0	S
0	8	0	11	Y
6	13	0	0	C
19	0	0	0	A
0	4	6	9	K
5	4	2	8	N

Name: CEBP_new
Binding Matrix:

A	C	G	T	
0	10	0	38	**T**
0	0	0	48	**T**
11	0	26	11	**G**
0	0	8	40	**T**
0	0	48	0	**G**
2	23	0	23	**Y**
48	0	0	0	**A**
48	0	0	0	**A**

Name: CEBP_rest1
Binding Matrix:

A	C	G	T	
0	3	0	6	T
7	0	0	2	A
0	0	0	9	T
0	0	0	9	T
0	0	5	4	K
0	0	9	0	G
0	6	0	3	C
0	0	0	9	T

Name: CEBP_rest10
Binding Matrix:

A	C	G	T	
0	0	7	0	G
0	0	7	0	G
0	3	4	0	S
2	0	3	2	D
0	0	7	0	G
5	1	1	0	A
0	0	7	0	G
0	0	7	0	G
4	0	0	3	W
0	0	7	0	G

2005:

Name: CEBP_sub10
Binding Matrix:

A	C	G	T	
0	0	0	61	T
0	0	9	52	T
20	0	36	5	R
0	61	0	0	C
35	20	6	0	M
17	44	0	0	C
38	23	0	0	M
61	0	0	0	A

Name: CEBP_sub11
Binding Matrix:

A	C	G	T	
29	4	6	1	A
0	0	0	40	T
0	2	6	32	T
2	0	19	19	K
3	26	10	1	C
1	17	7	15	Y
0	0	0	40	T
15	21	0	4	M
39	1	0	0	A
2	7	17	14	K

Name: CEBP_sub12
Binding Matrix:

A	C	G	T	
6	0	8	10	D
0	0	24	0	G
0	16	8	0	C
21	1	0	2	A
3	11	9	1	S
22	0	1	1	A
0	0	20	4	G
4	1	17	2	G

Name: CEBP_sub13
Binding Matrix:

A	C	G	T	
3	0	1	5	W
1	2	2	4	N
2	0	0	7	T
0	1	8	0	G
8	0	0	1	A
0	0	1	8	T
0	0	0	9	T
0	0	7	2	G
0	7	0	2	C
0	0	0	9	T

B. Re-identification of C/EBPTFBS by the 4 new matrices

No	AccNo	sub10	sub11	sub12	sub13
1	R03234	+	+	+	
2	Z_004	+	+	+	
3	Z_014	+	+	+	
4	Z_019	+	+	+	
5	Z_036	+	+	+	
6	R04487	+	+	+	
7	R08089	+	+	+	
8	R08091	+	+	+	
9	R14528	+	+	+	
10	Z_038	+	+		+
11	R02457	+	+		
12	R03128	+	+		
13	Z$U_001	+	+		
14	Z_002	+	+		
15	Z_018	+	+		
16	Z_037	+	+		
17	ZZR_001	+	+		
18	ZZR_001	+	+		
19	R00089	+	+		
20	R00102	+	+		
21	R00104	+	+		
22	R00111	+	+		
23	R00480	+	+		
24	R00629	+	+		
25	R00632	+	+		
26	R00837	+	+		
27	R00840	+	+		
28	R00847	+	+		
29	R01463	+	+		
30	R01686	+	+		
31	R01688	+	+		
32	R01689	+	+		
33	R01700	+	+		
34	R02045	+	+		
35	R02460	+	+		
36	R02468	+	+		
37	R02739	+	+		
38	R02895	+	+		
39	R02950	+	+		
40	R03141	+	+		
41	R03142	+	+		
42	R03143	+	+		
43	R03152	+	+		
44	R04480	+	+		
45	R04503	+	+		
46	R05074	+	+		
47	R08098	+	+		
48	R08130	+	+		
49	R08136	+	+		
50	R08878	+	+		
51	Z_005	+		+	
52	Z_043	+		+	
53	R00422	+		+	
54	R01456	+		+	
55	R02042	+		+	
56	R02461	+		+	
57	R03135	+		+	
58	R03322	+		+	
59	R03414	+		+	
60	R03733	+		+	
61	R04494	+		+	
62	R04514	+		+	
63	R08105	+		+	
64	R13167	+		+	

	AccNo	sub10	sub11	sub12	sub13
65	R14439	+		+	
66	R02458	+			
67	R04491	+			
68	Z_020	+			
69	Z_032	+			
70	Z_035	+			
71	R00082	+			
72	R00097	+			
73	R00156	+			
74	R00238	+			
75	R00239	+			
76	R00600	+			
77	R01156	+			
78	R01235	+			
79	R01344	+			
80	R01581	+			
81	R01587	+			
82	R01613	+			
83	R01839	+			
84	R02564	+			
85	R02583	+			
86	R02908	+			
87	R03069	+			
88	R03136	+			
89	R03148	+			
90	R03149	+			
91	R03730	+			
92	R08103	+			
93	R08123	+			
94	R08135	+			
95	R08138	+			
96	R13168	+			
97	R13169	+			
98	R13170	+			
99	R13171	+			
100	R13309	+			
101	R13567	+			
102	R14479	+			
103	R14487	+			
104	R14529	+			
105	R14533	+			
106	R14609	+			
107	Z_031	+			
108	R00637		+	+	
109	R01154		+	+	
110	R03242		+	+	
111	R03666		+	+	
112	R00103		+	+	
113	R01462		+	+	
114	R03250		+	+	
115	R08092		+	+	
116	R08099		+	+	
117	R13311		+	+	
118	R15892		+	+	
119	R03243		+		
120	Z_021		+		
121	R01184		+		
122	R02046		+		
123	R04477		+		+
124	R04505		+		
125	R04516		+		+
126	R08087		+		
127	R00093		+		
128	R00138		+		

129	R00464		+		
130	R00627		+		
131	R00698		+		
132	R00880		+		
133	R01217		+		
134	R01314		+		
135	R01358		+		
136	R01841		+		
137	R02744		+		
138	R03239		+		
139	R03251		+		
140	R04044		+		
141	R04047		+		
142	R04255		+		
143	R04476		+		
144	R04501		+		
145	R04512		+		
146	R04515		+		
147	R08086		+		
148	R08096		+		
149	R08116		+		
150	R13310		+		
151	R13463		+		
152	R00631			+	
153	R00635			+	
154	R02056			+	
155	R02584			+	
156	R02737			+	
157	R02745			+	
158	R02746			+	
159	R03068			+	
160	R03129			+	
161	R03139			+	
162	R03147			+	
163	R03244			+	
164	R03252			+	
165	R03732			+	
166	R04485			+	
167	R04486			+	
168	R04507			+	
169	R04662			+	
170	R08112			+	
171	R13179			+	
172	R02855				+
173	R03130				+
174	R03237				+
175	R03246				+
176	R08090				+
177	R13172				+
178	R00083				
179	R00090				
180	R01185				
181	R01445				
182	R02459				
183	R02740				
184	R02743				
185	R03131				
186	R03151				
187	R03307				
188	R03310				
189	R03657				
190	R04511				
191	R08097				
192	R08100				
193	R08113				

APPENDIX 2. Distance distributions

Estimation of the normality of distribution of $f_{d,\delta}$ and of the standart deviation.

To estimate the error of the predictions, we undertook computer simulations that showed that in the case when $M_A \ll L$ and $M_B \ll L$ the distribution of $f_{d,\delta}$ is close to normal and has the standard deviation equal to $\sigma \approx \sqrt{f_{d,\delta}/N}$.

We illustrate this result by an example where $M_A = 2$, $M_B = 3$, $L = 20$, $\delta = 1$. We generated 100 sets of 30 sequences each ($N = 30$), which contain M_A sites A and M_B sites B at random positions in the segment with the length L. The average numbers of pairs (per sequence) $f_{d,\delta}$ were directly measured (Fig.A1, red line). It was compared with the theoretical results $f^T_{d,\delta}$ (Fig.A1, blue line). Values of $f_{d,\delta}$ were different in different sets of sequences. This is shown on the plot in Fig.A2 for the case when $d = 0$; the blue line shows the theoretical result ($f^T_{d,\delta}$), the red line is the dependency of $f_{d,\delta}$ on the number n of set.

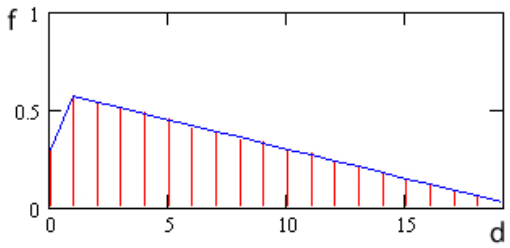


Figure A1. The average number of pairs (per sequence) $f_{d,\delta}$: Comparison of the theoretical results (blue line) with the measurements in a set of 100 random sequences (red line).

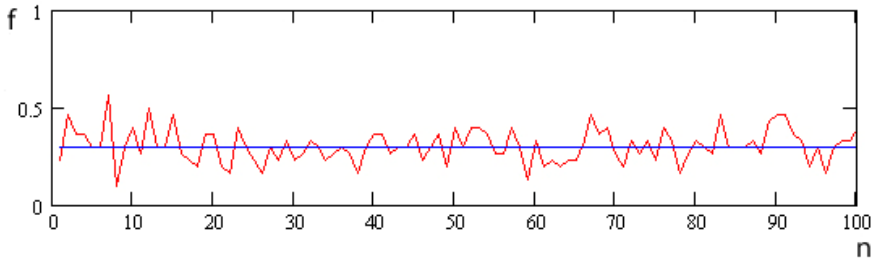


Figure A2. Dependence of the values of $f_{d,\delta}$ on the number of the set (n). Blue line – theoretical result, red line - measured in random sequences.

We analyzed the dependency of Δ on the number n, where Δ is the difference $f_{d,\delta}$ - $f^T_{d,\delta}$ normalized by the theoretical standart distribution $\sigma_T = \sqrt{f^T_{d,\delta}/N}$. The plot in Fig. A3 presents

the dependency of number of sets N_Δ on the Δ in the case $d = 0$, where blue line is the normal distribution with $\sigma = \sigma_T$.

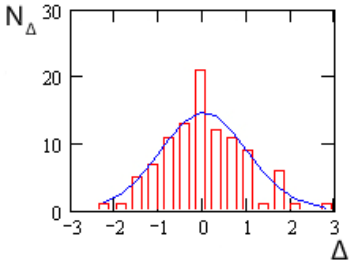


Figure A3. Dependence of number of sets N_Δ on the $\Delta = f_{d,\delta} - f^T_{d,\delta}$ in the case $d = 0$; blue line is the normal distribution with $\sigma = \sigma_{Teoretical}$.

We estimated the coincidence of histogram with normal distribution by calculation of the χ^2 parameter. The calculated values of χ^2 and critical values of χ^2 with the confidence equal to 0.95 for all d are shown in Table A1. It is clear that most of the N_Δ distributions are normal.

d	χ^2	χ^2_{crit}
0	5.64	12.59
1	11.52	15.51
2	6.54	15.51
3	11.56	14.07
4	9.66	14.07
5	4.7	14.07
6	9.32	14.07
7	1.99	14.07
8	16.63	14.07
9	3.59	14.07
10	8.91	12.59
11	12.46	14.07
12	6.14	11.07
13	9.02	11.07
14	16.34	9.49
15	1.45	9.49
16	3.97	7.81
17	3.55	7.81
18	4.17	5.99
19	10.95	3.84

Table A1. Calculated values of χ^2 and critical values of χ^2 with the confidence equal to 0.95 for all d.

We calculated the values of the standard deviations σ_n for each set n. The results are shown in Fig. A4.

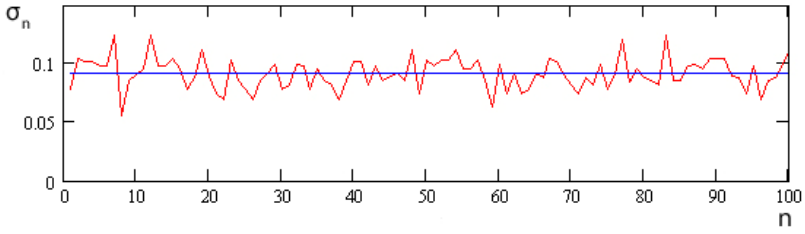


Figure A4. Dependence of σ_n on n in the case $d = 0$ (the blue line is σ_T).

We estimated the confidence bounds of the standard deviations (Mathematical statistics: basic ideas and selected topics. Peter J. Bickel, Kjell A. Doksum - 2nd ed., Prentice Hall, New Jersey, 2001).

Since the numbers of N in our experimental sets are close to 10, we get that $0.75\sigma_T \le \sigma \le 1.5\sigma_T$ with the confidence equal to 0.9.

To use the habitual criterion of 3 sigma we applied a renormalized quantity $4/3\sigma_T$. Therefore, as the analog of 3 standard deviations test, we used 4÷5 standard deviations test.

APPENDIX 3. P.a. promoter model

A. Search for 1 pair.

The numbers in the 3d and 4th column denote:
1 – AP-1
2 – C/EBP
3 – Elk-1
5– NF-κB
6 – Sp1

Denotations for pairs:
12- AP-1 – C/EBP, etc.

parameters:
Percentage:
100/80/40 (seed/whole TPset/Control):

whole set:

from	to	TF 1	TF 2	Pair class	% in (+)- training set	% in (-)- training set
3	83	1	5	1	82	39

Leave-1-out ("leave-2-out", because the 2 orthologs come together):
-1, 2:

from	to	TF 1	TF 2	Pair class	% in (+)- training set	% in (-)- training set
3	83	1	5	1	82	39

-3, 4: 100/70/40:

from	to	TF 1	TF 2	Pair class	% in (+)- training set	% in (-)- training set
10	77	1	5	1	76	37
10	83	1	5	1	79	38
43	129	1	5	1	73	39

-5, 6:

from	to	TF 1	TF 2	Pair class	% in (+)- training set	% in (-)- training set
3	77	1	5	1	79	39

-13, 14:

from	to	TF 1	TF 2	Pair class	% in (+)- training set	% in (-)- training set
3	83	1	5	1	82	39

-23, 24:

from	to	TF 1	TF 2	Pair class	% in (+)- training set	% in (-)- training set
3	83	1	5	1	82	39

-31, 32:

from	to	TF 1	TF 2	Pair class	% in (+)- training set	% in (-)- training set
3	83	1	5	1	82	39

+19, 20, -any pair:

from	to	TF 1	TF 2	Pair class	% in (+)- training set	% in (-)- training set
3	83	1	5	1	82	39

Conclusion:
AP-1 – NFκB (3-83), 1 fits as the only common pair.
The distance: to be adjusted.

Adjusting the distance:
Can be cut to 10 and prolonged to 93 without changes.

Result:
AP-1 – NFκB (10-93), 1

Search for 2 pairs does not give new pairs (*Suppmlementary materials*, "Search for 2 pairs_Pa_model.doc")

B. Complementary pairs:
Leave-1-out

Without distances:
Pairs left:

Pair1	orient.	Pair2	orient.
13	3	36	1
13	3	26	2
36	2	51	1
36	1	51	1
31	1	35	2
13	3	26	1
23	1	35	2
35	2	35	2
26	2	36	2
13	3	36	2
26	2	65	1
15	1	35	2
16	3	35	2
15	2	26	1
15	2	36	2
26	2	61	1
26	2	51	1
16	3	56	2
26	2	56	3
16	3	53	1
21	1	53	1
16	3	65	1
16	1	51	1
12	3	56	3

Checking with distances:

combinations left:

Pair1	Pair2
13 25 31 51 56	36 26 12 35

Any of the left can be combined with any of the right.

The best results:

$$\text{pairs\_or}_1 := \begin{pmatrix} 25 & 26 \\ 1 & 2 \\ 4 & 22 \\ 97 & 87 \end{pmatrix} \quad \text{pairs\_or}_2 := \begin{pmatrix} 13 & 36 \\ 3 & 1 \\ 28 & 14 \\ 39 & 96 \end{pmatrix} \quad \text{pairs\_or}_3 := \begin{pmatrix} 12 & 56 \\ 3 & 1 \\ 67 & 86 \\ 112 & 219 \end{pmatrix} \quad \text{pairs\_or}_4 := \begin{pmatrix} 13 & 26 \\ 3 & 2 \\ 28 & 190 \\ 39 & 219 \end{pmatrix}$$

$$\text{pairs\_or}_5 := \begin{pmatrix} 51 & 26 \\ 1 & 2 \\ 142 & 184 \\ 205 & 219 \end{pmatrix} \quad \text{pairs\_or}_6 := \begin{pmatrix} 25 & 12 \\ 1 & 3 \\ 4 & 103 \\ 97 & 112 \end{pmatrix}$$

That will be overfitted.

Finally:

$$S(m_1) = B^{(2)}_{C/EBP,Sp1}(22,87) \cup B^{(1)}_{C/EPB,NF-kB}(4,97)$$

$$S(m_2) = B^{(1)}_{Elk-1,Sp1}(14,96) \cup B^{(3)}_{AP-1,Elk-1}(28,39)$$

$$S(m_2) = B^{(3)}_{AP-1,C/EBP}(67,112) \cup B^{(1)}_{NF-kB,Sp1}(86,219)$$

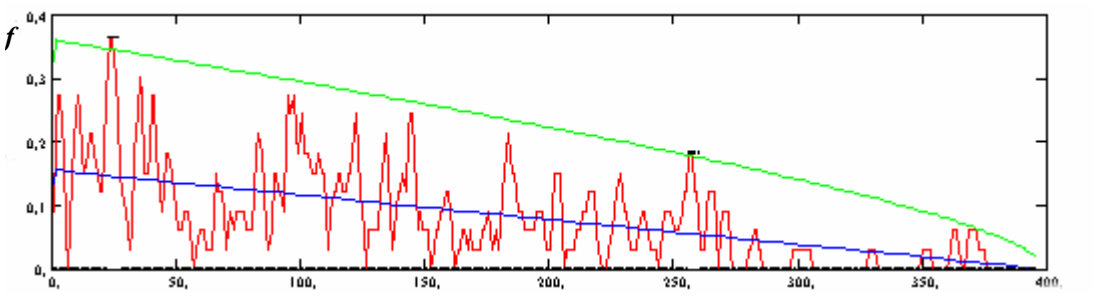
$$S(m_3) = B^{(2)}_{NF-\kappa B,Elk-1}(11,124) \cup B^{(3)}_{Ap-1,Elk-1}(28,39)$$

C. Re-identification of common and complementary pairs of P.a.-model by distance distributions approach

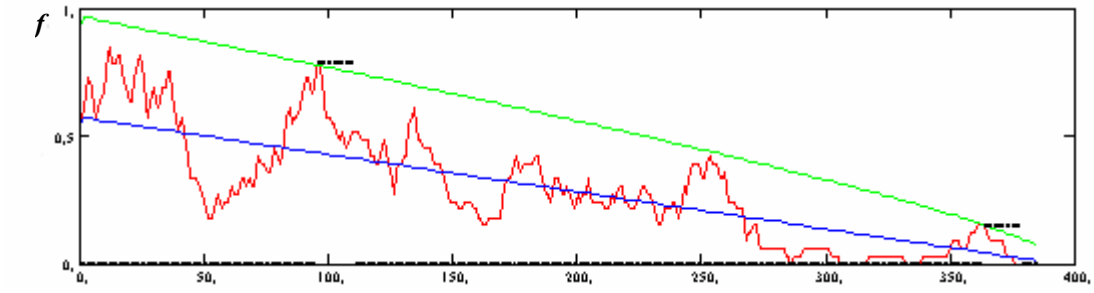
The method of distance distributions re-identifies the pairs selected for the previously made promoter model of the response of human epithelial cells to the binding of *P.aeruginosa*.
The pairs used in the model are marked with grey.
Only two pairs of the model (AP-1 - Elk-1 and NF- κB – Elk-1) have not been found.

	AP-1	Sp1	NF-κB	Elk-1	C/EBP
AP-1			+		+
Sp1			+	+	+
NF-κB					+
Elk-1					+
C/EBP					

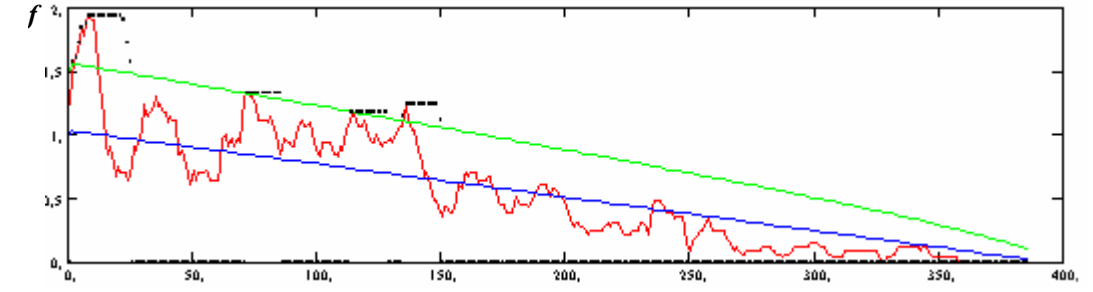
C/EBP-NF-κB: delta=4



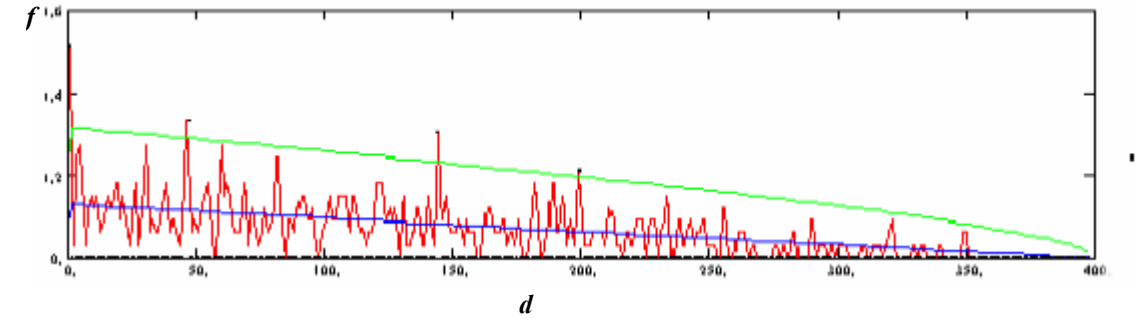
delta=15



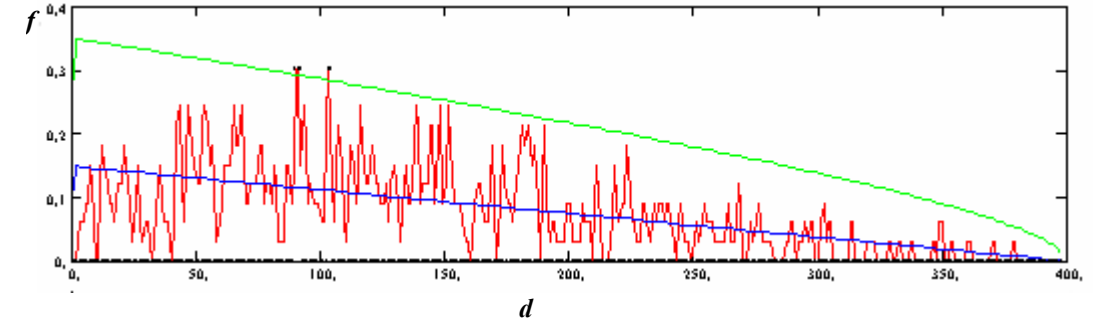
AP-1 – NF-κB, Delta=14



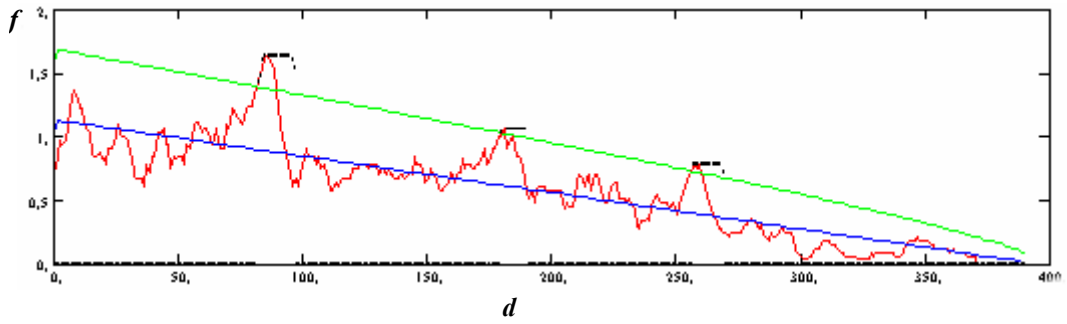
AP-1- C/EBP, Delta=5



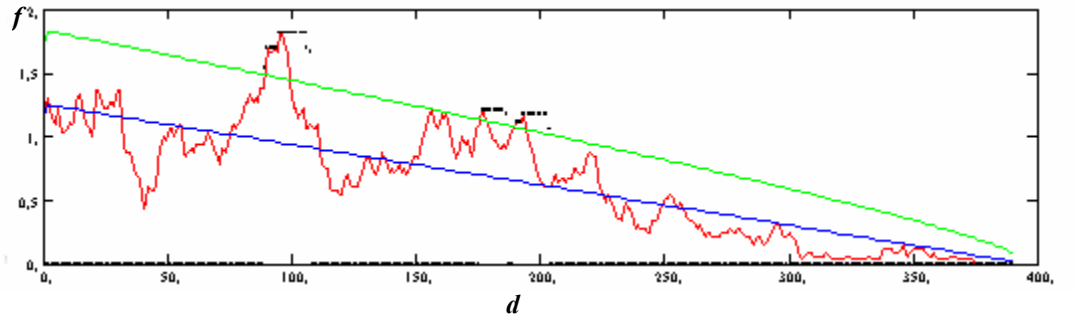
C/EBP-Sp1, Delta=3



Elk-1 – Sp1, Delta=10



NF-κB – Sp1, Delta=10



APPENDIX 4. LPS promoter model

Search for common pairs (LPS_90min)

$$pairs\_or_1 := \begin{pmatrix} 41 & 41 \\ 2 & 1 \\ 10 & 10 \\ 93 & 98 \end{pmatrix}$$

74% of the positive tzraining set.

Distances:
Sigma=5, delta=1,...25.

	AP-1	Sp1	CREB	NFAT	ATF-2	SRF	ETS	Elk-1	C/EBP	NF-κB
AP-1		+		+						+
Sp1			+			+				+
CREB				+		+				+
NFAT						+				
ATF-2										
SRF										+
ETS										
Elk-1										+
C/EBP										

Ddif=

	0	1	2	3	4	5	6	7	8	9
0	0	[48,3]	0	[11,3]	0	0	0	0	0	[23,3]
1	0	0	[41,3]	0	0	[3,3]	0	0	0	[24,3]
2	0	0	0	[6,3]	0	[2,3]	0	0	0	[8,3]
3	0	0	0	0	0	[3,3]	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	[1,3]
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	[1,3]
8	0	0	0	0	0	0	0	0	0	0

TF0=

	0	1
0	"AP-1"	91
1	"Sp1"	91
2	CREB"	91
3	"NFAT"	90
4	"ATF2"	90
5	"SRF"	89
6	"ETS"	89
7	"Elk-1"	89
8	C/EBP"	88
9	"NFkB"	81

	distances	% in TP
AP-1 - Sp1	26-78	88
AP-1 - NFAT	8-34	87
AP-1 - NF-κB	8-24 81-112	53 76 (8-112)
Sp1 - CREB	31-76	67
Sp1 - SRF	90-98	37
Sp1 - NF-κB	8-33	46
CREB - NF-κB	8-23	26
CREB - NFAT	9-23	67
CREB - SRF	56-62	24
NFAT - SRF	81-87	35
SRF - NF-κB	8-11	16
Elk-1 - NF-κB	33-45	25

Marked with grey - >50%

Denotations for TFs:
1 – AP-1
2 – ETS
3 – Elk-1
4 – NF-κB
5 – ATF2
6 – C/EBP
7 – CREB
8 – NFAT
9 – Sp1
10 - SRF

Denotations for pairs:
12- AP-1 – ETS, etc.
…..

Combinations:

$$\text{pairs\_or}_1 := \begin{pmatrix} 91 & 19 \\ 1 & 1 \\ 26 & 26 \\ 78 & 78 \end{pmatrix} \text{- 91\%}$$

$$+ \begin{pmatrix} 14 & 41 \\ 2 & 1 \\ 8 & 8 \\ 112 & 112 \end{pmatrix} \text{- 73\%}$$

$$\text{pairs\_or}_3 := \begin{pmatrix} 18 & 18 \\ 1 & 3 \\ 8 & 8 \\ 34 & 34 \end{pmatrix} \text{- 76\%}$$

$$+ \quad \text{pairs\_or}_2 := \begin{pmatrix} 97 & 79 \\ 1 & 3 \\ 31 & 31 \\ 76 & 76 \end{pmatrix} \text{- 70\%}$$

19+18+14=64%. Gives 19% control.

Without orientation:

$$\text{pairs\_and} := \begin{pmatrix} 18 & 0 & 78 & 0 & 0 & 19 \\ 8 & 31 & 9 & 8 & 46 & 26 \\ 34 & 76 & 23 & 112 & 136 & 78 \end{pmatrix} \quad \text{pairs\_or}_1 := \begin{pmatrix} 14 & 41 \\ 1 & 3 \\ 8 & 8 \\ 112 & 112 \end{pmatrix}$$

51% TP, 8,6%Control.
LPS_RT – 59% TP

$$\text{pairs\_or}_1 := \begin{pmatrix} 41 & 14 \\ 3 & 1 \\ 8 & 8 \\ 112 & 112 \end{pmatrix}$$

19(26 – 78) + =75%

19+78=63%
19+18=77%
19+18+14(or)=67%

again:

$$\text{pairs\_or}_1 := \begin{pmatrix} 91 & 19 \\ 1 & 1 \\ 26 & 26 \\ 78 & 78 \end{pmatrix}$$ - 90%

$$\text{pairs\_or}_1 := \begin{pmatrix} 91 & 19 \\ 1 & 1 \\ 26 & 26 \\ 78 & 78 \end{pmatrix} \quad \text{pairs\_or}_2 := \begin{pmatrix} 41 & 14 \\ 3 & 1 \\ 8 & 8 \\ 112 & 112 \end{pmatrix}$$ + - 77%

19+14+79=61%
....+18(and)=55%, 11%control

---+18+78+19(and)=5,6%control

	distances	single % in TP
AP-1 - Sp1	26-78	88
AP-1 - NFAT	8-34	87
AP-1 - NF-κB	8-24 81-112	53 76 (8-112)
Sp1 - CREB	31-76	67
CREB - NFAT	9-23	67

Triple combinations:

$$\text{triplets\_or}_1 := \begin{pmatrix} 40901 & 10904 & 0 \\ 8 & 26 & 0 \\ 112 & 78 & 140 \\ 26 & 8 & 0 \\ 78 & 112 & 140 \end{pmatrix}$$ 74%

$$\text{triplets\_or}_1 := \begin{pmatrix} 80109 & 10809 & 0 \\ 8 & 8 & 0 \\ 34 & 34 & 140 \\ 26 & 8 & 0 \\ 78 & 96 & 140 \end{pmatrix}$$

74%
both =57%, 17%control

$$\text{triplets\_or}_1 := \begin{pmatrix} 80109 & 10809 & 0 \\ 8 & 8 & 0 \\ 34 & 34 & 140 \\ 26 & 8 & 0 \\ 78 & 96 & 140 \end{pmatrix} \quad \text{triplets\_or}_2 := \begin{pmatrix} 40901 & 10904 & 0 \\ 8 & 26 & 0 \\ 112 & 130 & 140 \\ 26 & 8 & 0 \\ 130 & 112 & 140 \end{pmatrix}$$

$$\text{triplets\_or}_3 := \begin{pmatrix} 40908 & 80904 & 0 \\ 8 & 26 & 0 \\ 112 & 120 & 140 \\ 26 & 8 & 0 \\ 130 & 112 & 140 \end{pmatrix}$$

all three=53%, 13% control

$$\text{pairs\_or}_1 := \begin{pmatrix} 907 & 709 \\ 1 & 3 \\ 31 & 31 \\ 76 & 76 \end{pmatrix} \quad \text{triplets\_or}_1 := \begin{pmatrix} 40901 & 10904 & 0 \\ 8 & 26 & 0 \\ 112 & 130 & 140 \\ 26 & 8 & 0 \\ 130 & 112 & 140 \end{pmatrix}$$

63% TP, 21% control

$$\text{triplets\_or}_2 := \begin{pmatrix} 80109 & 10809 & 0 \\ 8 & 8 & 0 \\ 34 & 34 & 140 \\ 26 & 8 & 0 \\ 78 & 96 & 140 \end{pmatrix}$$

same + - 52%, 12% control.

Complementary pairs

$$\text{select1} := \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 & 30 & 31 & 32 & 33 & 34 & 35 & 36 & 37 & 38 & 39 & 40 & 41 & 42 & 43 & 44 & 45 & 46 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\text{select2} := \begin{pmatrix} 47 & 48 & 49 & 50 & 51 & 52 & 53 & 54 & 55 & 56 & 57 & 58 & 59 & 60 & 61 & 62 & 63 & 64 & 65 & 66 & 67 & 68 & 69 & 70 & 71 & 72 & 73 & 74 & 75 & 76 & 77 & 78 & 79 & 80 & 81 & 82 & 83 & 84 & 85 & 86 & 87 & 88 & 89 & 90 & 91 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

intersection=0,3 (intersection between 2 complementary subsets)
min=0,3 (minimal coverage of the whole set by one subset)
control=0,3 - nothing found

only with >2,5 fold induction

$$\text{select1} := \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 & 30 & 31 & 32 & 33 & 34 & 35 & 36 & 37 & 38 & 39 & 40 & 41 & 42 & 43 & 44 & 45 & 46 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\text{select2} := \begin{pmatrix} 47 & 48 & 49 & 50 & 51 & 52 & 53 & 54 & 55 & 56 & 57 & 58 & 59 & 60 & 61 & 62 & 63 & 64 & 65 & 66 & 67 & 68 & 69 & 70 & 71 & 72 & 73 & 74 & 75 & 76 & 77 & 78 & 79 & 80 & 81 & 82 & 83 & 84 & 85 & 86 & 87 & 88 & 89 & 90 & 91 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

APPENDIX 5. MALP-IRF promoter model

5a. Search for complementary pairs in the MALP-2/IRF-responsive set:

Columns:

1. minimal distance
2. maximal distance
3. pair
4. orientation
5. fraction in the corresponding subset
6. fraction in the complementary set
7. fraction in the control set

The numbers in the 3d column denote:
1 – AP-1
2 – ETS
3 – Elk-1
4 – NF-κB
5 – ATF2
6 – C/EBP
7 – CREB
8 – NFAT
9 – IRF
10 – SRF

Denotations for pairs:
12- AP-1 – ETS, etc.
…..
40 – Elk-1-SRF
80- CREB-SRF

Denotations for subsets:
F1 – subset 1
F2 – subset 2

Denotations for parameters:
Allowed overlap of complementary subsets:
"1in2=2" – insertion of subset 1 in subset 2 (here = 2)
"2in1=2" – insertion of subset 2 in subset 1 (here = 2)

0,8/0,9 – fraction of sequences containing a pair in correspondingly 1st (here 80%) and 2nd (here 90%) subsets.

Search 1: 1in2=3; 2in1=2; 0,8/0,7

F1 =

	0	1	2	3	4	5	6
0	8	109	40	1	0.82	0.38	0.49
1	8	133	67	1	0.82	0.38	0.48
2	3	85	71	1	0.82	0.38	0.44
3	3	97	71	1	0.84	0.38	0.46
4	3	98	71	1	0.86	0.38	0.47
5	4	98	71	1	0.82	0.38	0.43
6	8	95	72	1	0.82	0.38	0.4
7	8	102	72	1	0.84	0.38	0.41
8	15	102	72	1	0.82	0.38	0.39
9	6	59	80	1	0.82	0.25	0.41
10	6	62	80	1	0.84	0.25	0.42
11	8	62	80	1	0.82	0.25	0.42
12	23	85	80	2	0.82	0.25	0.43
13	23	106	80	2	0.84	0.38	0.48
14	60	147	80	2	0.82	0.38	0.46
15	27	89	80	3	0.82	0.25	0.45
16	27	110	80	3	0.84	0.38	0.5
17	27	116	80	3	0.86	0.38	0.51
18	27	132	80	3	0.89	0.38	0.53
19	28	110	80	3	0.82	0.38	0.49
20	30	111	80	3	0.82	0.38	0.49
21	30	116	80	3	0.84	0.38	0.5
22	30	132	80	3	0.86	0.38	0.52
23	40	116	80	3	0.82	0.38	0.46
24	40	132	80	3	0.84	0.38	0.49

F2 =

	0	1	2	3	4	5	6
0	66	126	19	1	0.75	0.02	0.06
1	66	138	19	1	0.88	0.05	0.06
2	112	138	19	1	0.75	0.05	0.03
3	26	69	19	2	0.75	0.05	0.05
4	68	128	19	2	0.75	0.05	0.06
5	68	140	19	2	0.88	0.05	0.07
6	125	140	19	2	0.75	0.02	0.02
7	58	108	29	2	0.75	0.05	0.04
8	26	99	29	3	0.75	0.05	0.06
9	24	43	49	1	0.75	0.05	0.02
10	24	53	49	1	0.88	0.05	0.02
11	34	53	49	1	0.75	0.05	0.01
12	43	60	49	1	0.75	0.02	0.01
13	46	97	49	1	0.75	0.02	0.03
14	26	55	49	2	0.75	0.05	0.02
15	26	56	49	2	0.88	0.05	0.02
16	36	56	49	2	0.75	0.05	0.02
17	45	62	49	2	0.75	0.02	0.01
18	50	99	49	2	0.75	0.02	0.02
19	23	50	49	3	0.75	0.05	0.02
20	23	76	49	3	0.88	0.05	0.03
21	60	140	59	2	0.75	0.05	0.08
22	47	80	69	2	0.75	0.05	0.05
23	8	52	79	2	0.75	0.02	0.05
24	4	4	89	1	0.88	0.02	0.03
25	10	44	89	1	0.75	0.05	0.06

1in2=3; 2in1=2; 0,8/**0,8**

F1 =

	0	1	2	3	4	5	6
0	8	109	40	1	0.82	0.38	0.49
1	8	133	67	1	0.82	0.38	0.48
2	3	85	71	1	0.82	0.38	0.44
3	3	97	71	1	0.84	0.38	0.46
4	3	98	71	1	0.86	0.38	0.47
5	4	98	71	1	0.82	0.38	0.43
6	8	95	72	1	0.82	0.38	0.4
7	8	102	72	1	0.84	0.38	0.41
8	15	102	72	1	0.82	0.38	0.39
9	6	59	80	1	0.82	0.25	0.41
10	6	62	80	1	0.84	0.25	0.42
11	8	62	80	1	0.82	0.25	0.42
12	23	85	80	2	0.82	0.25	0.43
13	23	106	80	2	0.84	0.38	0.48
14	60	147	80	2	0.82	0.38	0.46
15	27	89	80	3	0.82	0.25	0.45
16	27	110	80	3	0.84	0.38	0.5
17	27	116	80	3	0.86	0.38	0.51
18	27	132	80	3	0.89	0.38	0.53
19	28	110	80	3	0.82	0.38	0.49
20	30	111	80	3	0.82	0.38	0.49
21	30	116	80	3	0.84	0.38	0.5
22	30	132	80	3	0.86	0.38	0.52
23	40	116	80	3	0.82	0.38	0.46
24	40	132	80	3	0.84	0.38	0.49
25	46	132	80	3	0.82	0.38	0.47
26	9	98	107	1	0.82	0.38	0.52

F2=

	0	1	2	3	4	5	6
0	66	138	19	1	0.88	0.05	0.06
1	68	140	19	2	0.88	0.05	0.07
2	24	53	49	1	0.88	0.05	0.02
3	26	56	49	2	0.88	0.05	0.02
4	23	76	49	3	0.88	0.05	0.03
5	4	4	89	1	0.88	0.02	0.03
6	10	54	89	1	0.88	0.05	0.06
7	6	6	89	2	0.88	0.02	0.02
8	12	56	89	2	0.88	0.05	0.06
9	44	107	93	1	0.88	0.05	0.05
10	25	86	94	1	0.88	0.05	0.03
11	48	81	98	1	0.88	0.05	0.04

5b. Distance distributions

Over-represented distances in MALP-2-subset:

	AP-1	ETS	SRF	Elk-1	CREB	Sp1	ATF-2	NF-κB	NFAT	C/EBP
AP-1			+					+	+	
ETS								+		
SRF					+			+	+	
Elk-1								+		
CREB						+				
Sp1									+	
ATF-2										
NF-κB										
NFAT										+
C/EBP										

Pairs covering more than 50% TP:

	% of TP
SRF-CREB	60
SRF- NF-κB	70
Elk-1- NF-κB	67
SRF-NFAT	77
AP-1-NFAT	67

Pairs which can be combined:
(The numbers denote:
1 – AP-1
2 – ETS
3 – Elk-1
4 – NF-κB
5 – ATF2
6 – C/EBP
7 – CREB
8 – NFAT
9 – IRF
10 – SRF; the names of pairs should be read : 108=10-8="SRF – NFAT", etc.)

$$\text{pairs\_or}_1 := \begin{pmatrix} 108 & 108 \\ 1 & 3 \\ 5 & 55 \\ 34 & 80 \end{pmatrix} 77\%$$

$$\text{pairs\_or}_2 := \begin{pmatrix} 107 & 107 \\ 2 & 3 \\ 5 & 9 \\ 72 & 38 \end{pmatrix} - 71\%$$

+

$$\text{pairs\_or}_2 := \begin{pmatrix} 104 & 104 \\ 1 & 3 \\ 5 & 5 \\ 54 & 54 \end{pmatrix}$$ + - 64%

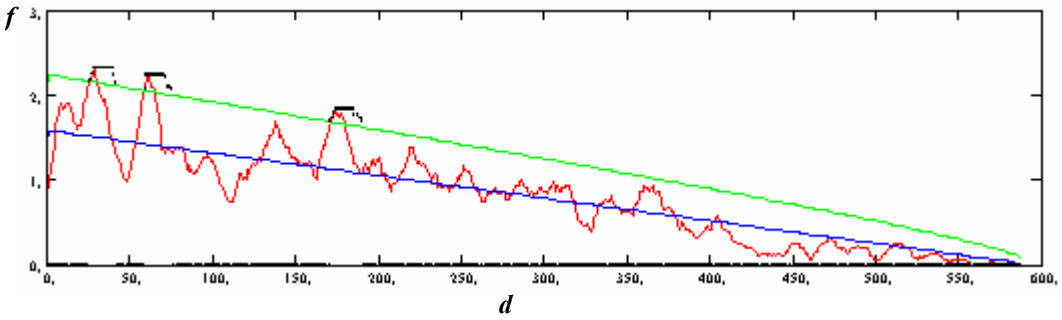
$$\text{pairs\_or}_2 := \begin{pmatrix} 34 & 34 \\ 1 & 3 \\ 18 & 18 \\ 49 & 49 \end{pmatrix}$$ + - 48%

$$\text{pairs\_or}_2 := \begin{pmatrix} 14 & 14 \\ 2 & 3 \\ 5 & 86 \\ 30 & 120 \end{pmatrix}$$ + - 40%

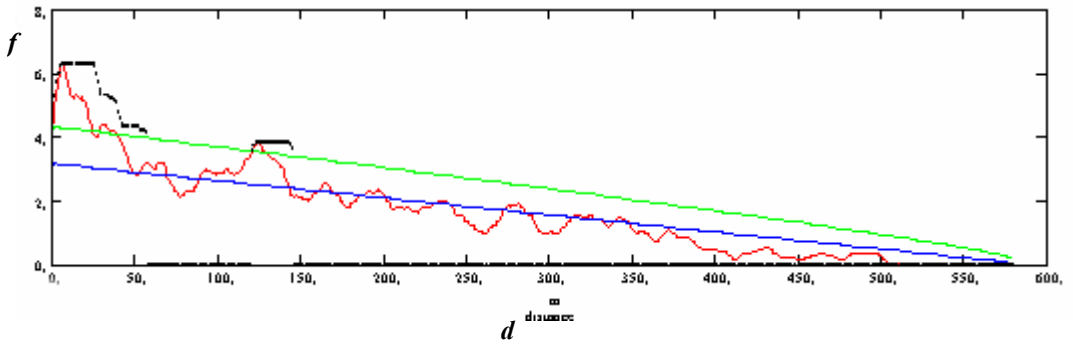
Accepted:
108+107=71%
108+104=64%

Result:
108+107+104=60% (12,4% Control).

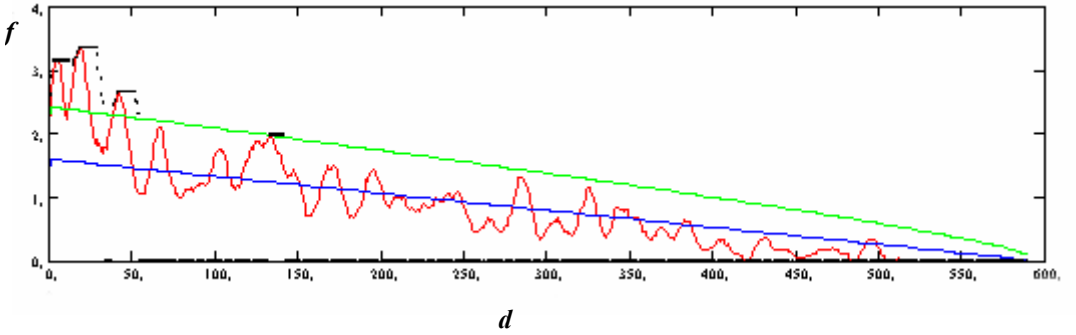
SRF-CREB, Delta=12, Sigma=4



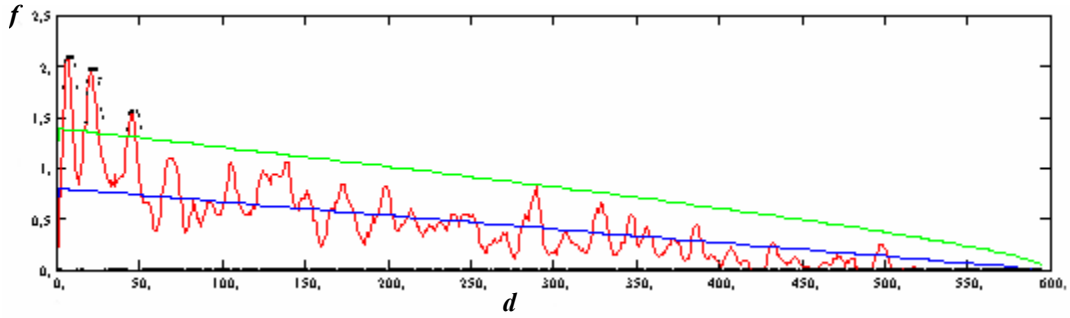
SRF-NFkB, Delta=20



delta=10

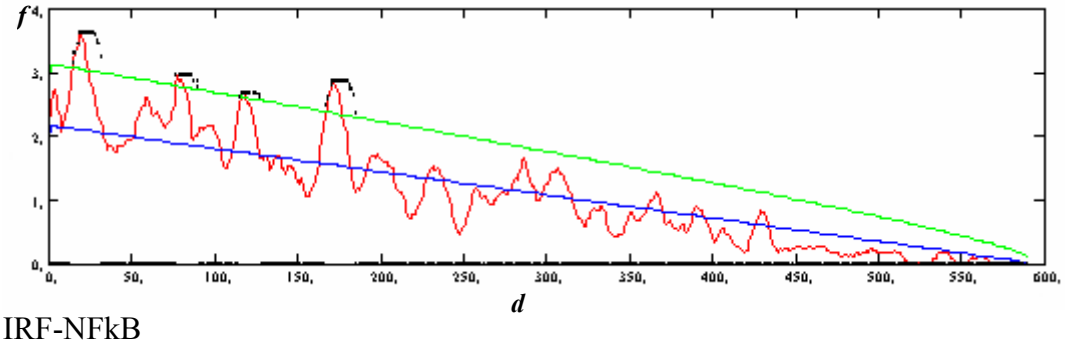


delta=5



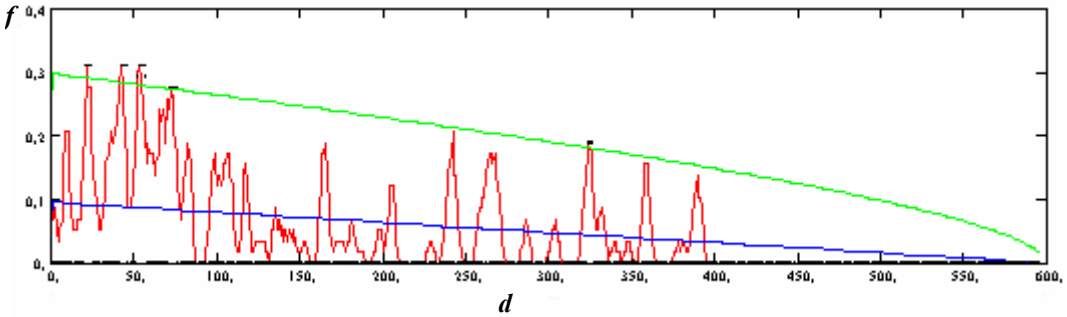
SRF-NFAT

Delta=10



IRF-NFkB

Delta=4



5c. Search for IRF-NFkB composite element (only IRF and NF-kB are considered)

Columns:

1. minimal distance
2. maximal distance
3. 1st TF
4. 2nd TF
5. pair class
6. fraction in the true positiveset
7. fraction in the control set

Denotations for TFs:
1 – NF-κB
2 - IRF

Seed=all IRF-responsive
min_select0=0.9
max_control=0.3

$$H = \begin{pmatrix} 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}$$

Leave-1-out

-1

$$H = \begin{pmatrix} 8 & 111 & 2 & 2 & 2 & 0.82 & 0.03 \\ 6 & 109 & 2 & 2 & 1 & 0.82 & 0.04 \\ 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}$$

-2

$$H = \begin{pmatrix} 8 & 111 & 2 & 2 & 2 & 0.82 & 0.03 \\ 6 & 109 & 2 & 2 & 1 & 0.82 & 0.04 \\ 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}$$

-3

$$H = \begin{pmatrix} 45 & 109 & 2 & 1 & 1 & 0.82 & 0.04 \\ 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 62 & 181 & 1 & 2 & 3 & 0.82 & 0.07 \\ 64 & 206 & 2 & 1 & 1 & 0.82 & 0.07 \\ 43 & 228 & 1 & 2 & 1 & 0.82 & 0.09 \\ 50 & 238 & 1 & 2 & 2 & 0.82 & 0.09 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}$$

-4:

$$H = \begin{pmatrix} 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}$$

-5

$$H = \begin{pmatrix} 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}$$

-6

$$H = \begin{pmatrix} 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}$$

-7

-8

H =

	0	1	2	3	4	5	6
0	23	56	1	2	3	0.82	0.03
1	25	58	2	1	1	0.82	0.03
2	45	109	2	1	1	0.82	0.04
3	23	99	1	2	3	0.91	0.05
4	25	109	2	1	1	0.91	0.06
5	62	181	1	2	3	0.82	0.07
6	64	206	2	1	1	0.82	0.07
7	24	228	1	2	1	0.91	0.1
8	26	238	1	2	2	0.91	0.1
9	2	2	2	2	2	1	0.15

$$H = \begin{pmatrix} 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}.$$

-9:

$$H = \begin{pmatrix} 24 & 53 & 1 & 2 & 1 & 0.82 & 0.03 \\ 26 & 56 & 1 & 2 & 2 & 0.82 & 0.03 \\ 23 & 56 & 1 & 2 & 3 & 0.82 & 0.03 \\ 25 & 58 & 2 & 1 & 1 & 0.82 & 0.03 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}.$$

-10:

$$H = \begin{pmatrix} 24 & 53 & 1 & 2 & 1 & 0.82 & 0.03 \\ 26 & 56 & 1 & 2 & 2 & 0.82 & 0.03 \\ 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 43 & 228 & 1 & 2 & 1 & 0.82 & 0.09 \\ 50 & 238 & 1 & 2 & 2 & 0.82 & 0.09 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}.$$

-11:

$$H = \begin{pmatrix} 23 & 99 & 1 & 2 & 3 & 0.91 & 0.05 \\ 25 & 109 & 2 & 1 & 1 & 0.91 & 0.06 \\ 24 & 228 & 1 & 2 & 1 & 0.91 & 0.1 \\ 26 & 238 & 1 & 2 & 2 & 0.91 & 0.1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 0.15 \end{pmatrix}.$$

% in control for the search with 1 pair is too high (2,8% in the best case, with loosing 40% of TP).

Acknowledgements

My first and the deepest gratitude I would like to express to Prof. Dr. Edgar Wingender, without whom this work would not be possible. His openness to new ideas, readiness to listen to any opinion, encouragement and support were of the most importance for me. Taken together with the giant scientific experience, encyclopedic knowledge of anything concerning transcription factors and around them, all these made the work with him a great experience.

My special thanks goes to Prof. Dr. Dieter Jahn, who made it all possible, having become the mentor of this work and the first referee, and appreciated this interdisciplinary work. Discussions with him shed for me light on many interesting sides of microbiological investigations.

I am grateful to both professors for being my referees, for sharing their time and efforts to make this work as good as possible.

I would like to thank all my colleagues from Göttingen and from Biobase, who helped me and made the work and stay in Germany a pleasure.

In Göttingen, I am indebted to the whole group for fruitful discussions during our seminars, for the inspiring questions and interest. My special thanks is to Tilman Sauer and Martin Haubrock for discussions and sharing the room; to Torsten Schöps, for being the system administrator and being always helpful; to Dr. Holger Michael, who is such a nice companion for the train journeys; to Carmen Modrok and Doris Waldmann, because they are the best secretaries I have ever seen; and, of course, to Dr. Anatolij Potapov, for his sincere interest, readiness for discussions and wise advice.

In Biobase, I would like to thank Dr. Olga Kel-Margoulis and Dr. Alexander Kel for the inspiring discussions and being always helpful and interested; Dr.Volker Matys for sharing interests; Claudia Choi for moral support; Dr. Klaus Hornischer for the help with data sets and attempts to teach me programming; Dr. Birgit Lewicky-Potapov for the important discussions in the morning hours; Dr. Ingmar Reuter and Dr. Heiko Saxel for the technical assistance; Dr. Ines Liebich for the discussions and friendly help; Dmitry Chekmenev for the technical help, discussions and being always ready for a cup of tea; Susanne Thiele, Doreen Kelterer and Anja Diehl for being always helpful.

I am especially grateful to our collaborators. I would like to thank Prof. Dr. Claus Scheidereit and Dr. Daniel Krappmann (MDC Berlin) for supplying us with data on LPS-triggering and helpful explanations and discussions.

My very sincere and deep gratitude I want to express to Prof. Dr. Peter F. Mühlradt, who not only supplied me with the data about the MALP-2-triggered pathway, but was always supportive and encouraging, and had time for many enlightening discussions.

I also want to thank the Intergenomics Bioinformatics Competence Center (Braunschweig) for financing this work, as part of the grant from the German Federal Ministry of Education and Research (grant No. 031U110A).

I want to thank my friends – those who stayed in Russia and believed in me, and those new friends I've got in Germany: Lena and Dmitry, Birgit and Anatolij, Claudia, Volker, Holger, Alena. Without the friends' support the life (and work) is impossible.

And the final gratitude goes to my family, for their understanding, encouragement and love. I want to especially thank my mother, who always believes in me and supports me; and my husband, who helped me in all possible ways, being my best scientific adviser and supporter and taking care of everything in my life.